

Srija Unnikrishnan
Sunil Surve
Deepak Bhoir (Eds.)

Communications in Computer and Information Science

125

Advances in Computing, Communication and Control

International Conference, ICAC3 2011
Mumbai, India, January 2011
Proceedings

Srija Unnikrishnan Sunil Surve
Deepak Bhoir (Eds.)

Advances in Computing, Communication and Control

International Conference, ICAC3 2011
Mumbai, India, January 28-29, 2011
Proceedings

Volume Editors

Srija Unnikrishnan
Sunil Surve
Deepak Bhoir

Fr. Conceicao Rodrigues College of Engineering Fr. Agnel Ashram
Bandra (West) Mumbai, India

E-mail: {srija, surve, bhoir}@frcrce.ac.in

ISSN 1865-0929

e-ISSN 1865-0937

ISBN 978-3-642-18439-0

e-ISBN 978-3-642-18440-6

DOI 10.1007/978-3-642-18440-6

Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010942873

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

© Springer-Verlag Berlin Heidelberg 2011

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

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

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

Printed on acid-free paper

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

Preface

The International Conference on Advances in Computing, Communication and Control (ICAC3) is a biennial Conference held by the Fr. Conceicao Rodrigues College of Engineering, Mumbai, India. The second in the series, ICAC3 2011, was organized during January 28–29, 2011. ICAC3 2011 received 309 submissions from 5 countries and regions. After a series of rigorous reviews, 67 high-quality papers were selected for publication in the ICAC3 2011 proceedings.

The primary goal of the conference was to promote research and development activities in computing, communication and control in India and the rest of the world. ICAC3 2011 provided a forum for engineers and scientists in academia, industry, and government organizations to address the most innovative research and development and to present and discuss their ideas, results, and experiences on all aspects of the above-mentioned areas.

The papers were selected after a series of reviews; one non-blind review to check suitability of the paper for the conference followed by three blind reviews out of which two were by experts and the third by peer authors.

The fruitful outcome of the conference was the result of a well-coordinated effort of the Program and Technical Committee members as well as the Advisory Committee. We would like to express our heartfelt thanks to them for their expertise, time, and enthusiasm. We are also grateful to the members of the local Organizing Committee for supporting us in handling so many organizational tasks and to the keynote speakers for accepting our invitation. Also, we would like to thank Springer for supporting the publication of the proceedings. Last, but not the least, we would like to thank the authors for submitting their work and hope that the participants enjoyed the conference.

January 2011

Srija Unnikrishnan
Sunil Surve
Deepak Bhoir

Organization

ICAC3 is a biennial conference organized by Fr. Conceicao Rodrigues College of Engineering, Mumbai, India.

Patrons

Fr. F. Diniz, Superior

Fr. Victor, Director

Executive Committee

General Chair

Sunil Surve

Co-chair

Deepak Bhoir

Convenor

B.R. Prabhu

Organizing Chair

Merly Thomas

Shilpa Patil

Prachi Cheoolkar

Monica Khanore

Financial Chair

Track Chairs

Computing

Jagruti Save

Communication

Mahesh Sharma

Control

K. Narayanan

Advisory Committee

Vijay K. Kanabar

Metropolitan College Boston University,
Boston, USA

H.B. Kekre

NMIMS, Mumbai, India

Sunil Survaiya

Research and Development, Corning Cable
Systems, Germany

Sameer Hemmady

SAIC, USA

N.M. Singh

VJTI, Mumbai, India

Sudhir Sawarkar

Datta Meghe College of Engineering,
Mumbai, India

S.T. Gandhe

Sardar Patel Institute of Technology,
Mumbai, India

Program/Technical Committee

Program Chair Vijay K. Kanabar	Srija Unnikrishnan Metropolitan College Boston University, Boston, USA
Mudasser F. Wyne Sriman Narayan Iyengar	National University, USA School of Computing Science and Engineering, VIT University, Vellore, India
Maninder Singh Thapar Deepak Garg Thapar Satish R. Kolhe S. Sane Sudhir Sawarkar	Patiala University, India Punjab University, India North Maharashtra University, Jalgaon, India VJTI, Mumbai, India Datta Meghe College of Engineering, Mumbai, India
S. Nikolas	National Institute of Technology, Tiruchirapalli, India
Deven Shah	Sardar Patel Insitute of Technology, Mumbai, India
Tanuja Sarode	Thadomal Shahani Engineering College, Mumbai, India
Pradip Peter Dey	Department of Computer Science and Information Systems, National University San Diego, California
Avinash Keskar	Visvesvaraya National Institute of Technology, Nagpur, India
Archana Kale	Thadomal Shahani Engineering College, Mumbai, India
M. Kritika	Fr. C. Rodrigues Institute of Technology, Vashi, India
Radha Shankar Mani	Sardar Patel Insitute of Technology, Mumbai, India
Rohin Daruwala Jayant Gadge	VJTI, Mumbai, India Thadomal Shahani Engineering College, Mumbai, India
Uday Khedker Ketan Shah Madhuri Bhavser	IIT Bombay, India NMIMS Mumbai, India Nirma University of Science and Technology, Ahmedabad, India
Maniklal Das Mazad S. Zaveri Manoj Devare	DA-IICT Ganghinagar, India DA-IICT Ganghinagar, India Vidya Pratishthan's Institute of Information Technology, Baramati, India

S.T. Gandhe	Sardar Patel Institute of Technology, Mumbai, India
Maulika S. Patel	G.H. Patel College of Engineering and Technology, Vallabh Vidyanagar, India
T.J. Parvat	Sinhgad Institute of Technology, Lonavala, India
Santosh Bothe	Sinhgad Institute of Technology, Lonavala, India
S.V. Pingle	Sinhgad Institute of Technology, Lonavala, India
Nitin Dhawas	Sinhgad Institute of Technology, Lonavala, India
Abhinay Pandya	DA-IICT Ganghinagar, India
Jayant Umale	D.J. Sanghvi College of Engineering, Mumbai, India
Kamal Shah	St. Francis Institute of Technology Borivali, India
H.B. Kekre	NMIMS, Mumbai, India
S. SelvaKumar	National Institute of Technology, Tiruchirapalli, India
K.V.V. Murthy	Amrita Vishwa Vidyapeetham, Bengaluru, India
R.N. Awale	VJTI, Mumbai, India
P. Muthu Chidambaranathan	National Institute of Technology, Trichi, India
Mahesh Chandra	Birla Institute of Tehnology, Ranchi, India
Vivekanand Mishra	S.V. National Institute of Technology, Surat, India
Vijay Janyani	Malaviya National Institute of Technology, Jaipur, India
R.K. Shevgaonkar	IIT Bombay, India
B.K. Lande	VJTI, Mumbai, India
S.T. Hamde	SCGS, Nanded, India
Mahesh B. Kokare	SGGS, Nanded, India
Anjali Malviya	Thadomal Shahani Engineering College, Bandra, Mumbai, India
G.P. Bhole	VJTI, Mumbai, India
Suprava Patnaik	S.V. National Institute of Technology, Surat, India
Ashish Khare	Allahabad University, India
Sanjay Talbar	SGGS, Nanded, India
V.K. Govindan	National Institute of Technology, Calicut, India
Amol Khanapurkar	Tata Consultancy Services Ltd., India

Sincy George	Fr. CRIT, Vashi, Navi Mumbai, India
M.R. Vikraman	NSS College of Engineering, Palghat, India
M.V. Pitke	Nicheken Technologies, Chennai, India
Satya Sheel	Motilal Nehru National Institute of Technology, Allahabad, India
Sameer Hemmady	SAIC, USA
Akhilesh Swarup	Galgotias College of Engineering and Technology, Greater Noida, India
Taskeen Ameer Nadkar	IIT Bombay, India
Uday Pandit	Thadomal Shahani Engineering College, Bandra, Mumbai, India
Sandipan P Narote	Sinhgad College of Engineering, India
Dattatray V. Jadhav	TSSM (JSPM) Bhivarabai Sawant College of Engineering and Research, India
Lucy J. Gudino	BITS-GOA, India
Baidyanath Roy	Bengal Engineering and Science University, Shibpur, India
Chintan Modi	G. H. Patel College of Engineering and Technology, Vallabh Vidyanagar, India
Sameer	NIT-C, India
Sanjay Kumar	BIT-Mesra, India
Janardhan Sahay	BIT-Mesra, India
Upena D. Dalal	Sardhar Vallabhai National Institute of Technology, India
P. Muthu Chidambara Nathan	NIT-T, India
M.B. Kokare	Shri Guru Gobind Singhji Institute of Engineering and Technology, Vishnupuri Nanded, India
Kailash Karande	Sinhgad Institute of Technology, Lonavala, Pune, India
S.C. Sahasrabudhe	DAIICT, India
Manik Lal Das	DAIICT, India
T.J. Siddiqui	J.K. Institute of Applied Physics and Technology University of Allahabad, India
A.G. Patil	S.B.M. Polytechnic, Irla, Mumbai, India
J.D. Mallapur	BEC, Bagalkot, India
Mamata S. Kalas	KITS College of Engineering, Kolhapur, India
Amutha Jay Kumar	VJTI, Mumbai, India
Poornima Talwai	RAIT, Mumbai, India
B. K. Mishra	Thakur College of Engineering, Mumbai, India
Maniroja	Thodamal Sahani College of Engineering, Mumbai, India

Kishor S. Kinage	D.J. Sanghvi College of Engineering, Mumbai, India
K.T.V. Reddy	Terna Engineering College, Mumbai, India
R.A. Jadhav	K.J. Somayya College of Engineering, Mumbai, India
Kirti Panwar	School of Electronics, Devi Ahilya University, Indore, India
Nar Singh	J.K. Institute of Applied Physics and Technology, University of Allahabad, India
Pallavi Yarde	School of Electronics, Devi Ahilya University, Indore, India
Satish Ket	RGIT, Mumbai, India
Sukanya kulkarni	SPIT, Mumbai, India
Bharti Singh	K.J. Somayya College of Engineering, Mumbai, India
Devendra Singh Bais	School of Electronics Devi Ahilya University, Indore, India
Tanaji Dadarao Biradar	D.J. Sanghvi College of Engineering, Mumbai, India

Local Organizing Committee

Sapna Prabhu	Swapnali Mahdik
Roshni Padate	Ranjushree Pal
Darshana shah	Kalpna Devrukhkar
Swati Ringe	Sushma Nagdeote
Hemant Khanolkar	Ashwini Avad
Deepali Koshti	Parshvi Shah
Supriya Khamoji	Sarika Kumawat
Srikant Thati	Vipin Palkar
Ganesh Bhirud	Binsy Joseph
B.S. Daga	

Table of Contents

Computing

Modified Trivial Rejection Criteria in Cohen-Sutherland Line Clipping Algorithm	1
<i>Jeet Kumar and Ashish Awasthi</i>	
Greedy Views Selection Using Size and Query Frequency	11
<i>T.V. Vijay Kumar and Mohammad Haider</i>	
Quantum-Inspired Differential Evolution on Bloch Coordinates of Qubits	18
<i>Ankit Pat, Ashish Ranjan Hota, and Avneet Singh</i>	
Extending Speech-Act Based Communication to Enable Argumentation in Cognitive Agents	25
<i>Punam Bedi and Pooja Vashisth</i>	
Medical Diagnosis Using Generic Intelligent Agents	41
<i>Mukesh Kumar</i>	
Computational Modeling and Dynamical Analysis of Genetic Networks with FRBPN- Algorithm	49
<i>Raed I. Hamed</i>	
A Petri Net-Fuzzy Predication Approach for Confidence Value of Called Genetic Bases	56
<i>Raed I. Hamed</i>	
Textural Feature Based Image Classification Using Artificial Neural Network	62
<i>Salavi Rashmi and Sohani Mandar</i>	
Data-Warehousing Applications in Manufacturing Industry – Applicable Solutions and Challenges Faced	70
<i>Goparaju V. Ramesh, Sattiraju N. Rao, and Mogalla Shashi</i>	
Application of Clustering in Virtual Stock Market	79
<i>Kavita M. Gawande and Sangita C. Patil</i>	
Distribution of Loads and Setting of Distribution Sub Station Using Clustering Technique	88
<i>Shabbiruddin and Chakravorty Sandeep</i>	
Landscape of Web Search Results Clustering Algorithms	95
<i>Ujwala Bharambe and Archana Kale</i>	

An Improved K-Means Clustering Approach for Teaching Evaluation . . . 108
Oswal Sangita and Jagli Dhanamma

Consensus Based Dynamic Load Balancing for a Network of Heterogeneous Workstations 116
Janhavi Baikerikar, Sunil Surve, and Sapna Prabhu

An Effective Way to Hide the Secret Audio File Using High Frequency Manipulation 125
Mahendra Kumar Pandey, Girish Parmar, and Sanjay Patsariya

Web Usage Mining: An Implementation View 131
Sathya Babu Korra, Saroj Kumar Panigrahy, and Sanjay Kumar Jena

A Genetic Algorithm Way of Solving RWA Problem in All Optical WDM Networks 137
Ravi Sankar Barpanda, Ashok Kumar Turuk, Bibhudatta Sahoo, and Banshidhar Majhi

Working of Web Services Using BPEL Workflow in SOA 143
Aarti M. Karande, Vaibhav N. Chunekar, and B.B. Meshram

A Meta Search Approach to Find Similarity between Web Pages Using Different Similarity Measures 150
Jaskirat Singh and Mukesh Kumar

Cost Estimation for Distributed Systems Using Synthesized Use Case Point Model 161
Subhasis Dash and Arup Abhinna Acharya

Comparative Approach to Cloud Security Models 170
Temkar Rohini

Development of Agile Security Framework Using a Hybrid Technique for Requirements Elicitation 178
Sonia and Archana Singhal

Accuracy Comparison of Predictive Algorithms of Data Mining: Application in Education Sector 189
Mamta Sharma and Monali Mavani

Orchestrator Model for System Security 195
Aradhana Goutam, Rajkamal, and Maya Ingle

Communication

Performance Analysis of Interleave Division Multiple Access Scheme with Different Coding Techniques 200
Parul Awasthi, Sarita Singh Bhadauria, and Madhuri Mishra

Channel Assignment to Minimize Interference in Multiradio Wireless Mesh Networks	208
<i>S. Sekhar Babu and V. Sumalatha</i>	
IEC 61850: Goose Messaging Implementation for MPR	214
<i>Hemalata M. Shingate, Srijia Unnikrishnan, and Sudarshan Rao</i>	
Performance Analysis of Energy Efficient Routing Algorithms for Adhoc Network	222
<i>Dhiraj Nitnaware and Ajay Verma</i>	
ECG Signal Compression Using Different Techniques	231
<i>K. Ranjeet, A. Kumar, and R.K. Pandey</i>	
Efficient Content Based Video Retrieval for Animal Videos	242
<i>Vijay Katkar and Amit Barve</i>	
Data Randomization for Synchronization in OFDM System	249
<i>Rakhi Thakur and Kavita Khare</i>	
Split Step Method in the Analysis and Modeling of Optical Fiber Communication System	254
<i>Saroja V. Siddamal, R.M. Banakar, and B.C. Jinaga</i>	
Performance Analysis of MIMO- Space-Time Trellis Code System over Fading Channels	262
<i>Sandeep Bhad and A.S. Hiwale</i>	
Modeling Performance Evaluation of Reinforcement Learning Based Routing Algorithm for Scalable Non-cooperative Ad-hoc Environment	269
<i>Shrirang Ambaji Kulkarni and G. Raghavendra Rao</i>	
SRPV: A Speedy Routing Protocol for VANET	275
<i>Suparna DasGupta and Rituparna Chaki</i>	
Simultaneous Multiple Link/Node Failure Handling for Different Service-Paths in MPLS Networks	285
<i>Shah Rinku and Chatterjee Madhumita</i>	
Energy-Efficient Multilevel Clustering in Heterogeneous Wireless Sensor Networks	293
<i>Vivek Katiyar, Narottam Chand, and Surender Soni</i>	
Anomaly Detection in Ethernet Networks Using Self Organizing Maps	300
<i>Saroj Kumar Panigrahy, Jyoti Ranjan Mahapatra, Jignyanshu Mohanty, and Sanjay Kumar Jena</i>	

A Heuristic Multi Criteria Routing Protocol in Wireless Sensor Networks	306
<i>Alireza Shams Shafigh and Marjan Niyati</i>	
Framework and Implimentation of an Agent Based Congestion Control Technique for Mobile Ad-hoc Network	318
<i>Sarita Singh Bhadauria and Vishnu Kumar Sharma</i>	
A Hybrid Algorithm for Satellite Image Classification	328
<i>Samiksha Goel, Arpita Sharma, and V.K. Panchal</i>	
An Image Authentication Technique in Frequency Domain Using Secure Hash Algorithm (FDSHA)	335
<i>Amitava Nag, Debasish Biswas, Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, and Partha Pratim Sarkar</i>	
A Robust and Fast Text Extraction in Images and Video Frames	342
<i>Anubhav Kumar, Awanish Kr. Kaushik, R.L. Yadav, and Anuradha</i>	
A Chip-Based Watermarking Framework for Color Image Authentication for Secured Communication	349
<i>Soumik Das, Pradosh Banerjee, Monalisa Banerjee, and Atal Chaudhuri</i>	
An Image Decomposition Parallel Approach to Train Flann for an Adaptive Filter	355
<i>Manoj Kumar Mishra, Nachiketa Tarasia, Bivasa Ranjan Parida, and Sarita Das</i>	
Palette Based Technique for Image Steganography	364
<i>Anuradha Lamgunde and Achana Kale</i>	
Bilingual Malayalam – English OCR System Using Singular Values and Frequency Capture Approach	372
<i>Bindu A. Thomas and C.R. Venugopal</i>	
Minimization of Handoff Failure and Cellular Traffic by Introducing IEEE 802.11b WLAN Router in the Handoff Region	378
<i>Tapas Jana, Joydeep Banerjee, Indranil Chakroborty, Tara Sankar Patra, Debabrata Sarddar, M.K. Naskar, and Utpal Biswas</i>	
Design of GSM Based Auto-responder for Educational Institute	386
<i>D.D. Vyas and H.N. Pandya</i>	

Minimization of Handoff Latency by Area Comparison Method Using GPS Based Map	393
<i>Tapas Jana, Joydeep Banerjee, Subhojit Banerjee, Souvik Kumar Mitra, Debabrata Sarddar, M.K Naskar, and Utpal Biswas</i>	
Scalable Distributed Diagnosis Algorithm for Wireless Sensor Networks	400
<i>Arunanshu Mahapatro and Pabitra Mohan Khilar</i>	
Comparison of Routing Protocols for MANET and Performance Analysis of DSR Protocol	406
<i>Parma Nand and S.C. Sharma</i>	
Transmitter Based Capacity Enhancement with Cross-Layer Design Approach for IEEE 802.11 Ad-hoc Networks	413
<i>Satish Ket and R.N. Awale</i>	
Wireless Sensor Network Using Bluetooth	424
<i>Omkar Javeri and Amutha Jeyakumar</i>	
A Performance of Security Aspect in WiMAX Physical Layer with Different Modulation Schemes	433
<i>Rakesh Kumar Jha, Suresh Limkarl, and Upena D. Dalal</i>	
CE-OFDM: A PAPR Reduction Technique	441
<i>R.S. Chaudhari and A.M. Deshmukh</i>	
Initializing Cluster Center for K-Means Using Biogeography Based Optimization	448
<i>Vijay Kumar, Jitender Kumar Chhabra, and Dinesh Kumar</i>	
Implementation of Parallel Image Processing Using NVIDIA GPU Framework	457
<i>Brijmohan Daga, Avinash Bhute, and Ashok Ghatol</i>	
Control	
Toggle Coverage for ALU Using VHDL	465
<i>D. Venkat Reddy, Ch.D.V. Paradesi Rao, and E.G. Rajan</i>	
Design and Implementation of Voltage Control Oscillator (VCO) Using 180nm Technology	472
<i>M.R. Halesh, K.R. Rasane, and H. Rohini</i>	
Design of Robust PID Controller for Flexible Transmission System Using Quantitative Feedback Theory (QFT)	479
<i>Mukesh D. Patil and Kausar R. Kothawale</i>	

Flatness Based Formation Control of Non Holonomic Vehicle	486
<i>Ch. Venkatesh, Sunil K. Surve, and N.M. Singh</i>	
High Performance Tracking Controller for the Class of Uncertain Discrete-Time System with Input Delay	494
<i>Deepti Khimani and Machhindranath Patil</i>	
Implementation of Controller Area Network (CAN) Bus (Building Automation)	507
<i>S. Ashtekar Shweta, D. Patil Mukesh, and B. Nade Jagdish</i>	
DXCCII-Based Mixed-Mode Electronically Tunable Quadrature Oscillator with Grounded Capacitors	515
<i>Mohd. Samar Ansari and Sumit Sharma</i>	
Stereo Matching for 3D Building Reconstruction	522
<i>Gaurav Gupta, R. Balasubramanian, M.S. Rawat, R. Bhargava, and B. Gopala Krishna</i>	
Fingerprint Identification Using Sectionized Walsh Transform of Row and Column Mean	530
<i>H.B. Kekre, Tanuja K. Sarode, and Rekha Vig</i>	
Author Index	537

Modified Trivial Rejection Criteria in Cohen-Sutherland Line Clipping Algorithm

Jeet Kumar and Ashish Awasthi

Shri Ramswaroop Memorial Group of Professional Colleges,
Lucknow, Uttar Pradesh, India
jeetkm@sify.com, ashish3awasthi@rediffmail.com

Abstract. In the line clipping procedures if the line is not completely inside the clipping window then we have no option but to divide the line into segments at the intersections of the line and clipping window edges and then identify which segment is inside and which segment is outside the clipping window. But if the line is completely outside the clipping window, the line is eligible for total clipping or trivial rejection. There exist total 26 possible cases for trivial rejection of a line. Out of these 26 cases only 12 cases were solved without dividing the line into segments in the Cohen-Sutherland line clipping algorithm. Remaining 14 cases were solved by dividing the line into segments. But if the line is totally outside the clipping window, we don't require any segmentation. This paper presents the mechanism by which the remaining 14 cases can be solved without dividing the line into segments.

Keywords: Line-clipping, clipping-window, total-clipping, trivial-rejection and Cohen-Sutherland.

1 Introduction

Line clipping procedure is the mechanism that tells how to cut off the lines which are outside the clipping window [1], [2], [5]. The line clipping procedure can be broadly classified into three disjoint classes. Class-1 contains the lines which are completely inside the clipping window and therefore are trivially accepted. Class-2 contains the lines which are partially inside the clipping window, for these lines we have no options but to divide the lines into segments at the intersections of the line and clipping window edges and then examine each segment that whether the segment is completely inside or completely outside the clipping window. Class-3 contains the lines which are completely outside the clipping window, and therefore are trivially rejected. This paper focuses on the lines which are trivially rejected. The mechanism provided in this paper improves the performance of Cohen-Sutherland line clipping algorithm by modifying the trivial rejection criteria. Contrast to Cohen-Sutherland line clipping algorithm, our mechanism doesn't divide any line into segments if the line is completely outside the clipping window, therefore eliminates the overhead of segmentation of a line in more than half of the cases. All the cases of class-3, where the line is

completely outside the clipping window, are discussed in detail and proved that no segmentation of lines is required.

We have organized the contents of the paper by firstly describing the region code assignment procedure of Cohen-Sutherland line clipping algorithm in the literature review section. After this four procedures are defined that test a line for intersection at any of the clipping window edges. If the line intersects any of the clipping window edges, then this is the case of class-2 lines (partially inside) and therefore beyond the scope of this paper. If a line doesn't intersect any of the clipping window edges and doesn't fall under class-1 (completely inside), then the line is trivially rejected. In the next section 14 out of 26 cases of trivial rejection, for which bitwise AND of region codes of the end points of a line is zero, are discussed and solved without dividing the line into segments.

2 Literature Review

In the Cohen-Sutherland line clipping algorithm every line end-point is assigned a four-bit binary code called region code that identifies the location of the point relative to the boundaries of the clipping window rectangle [1], [2], [3], [4], [5], [6], [7], [8].

1001	1000	1010
0001	0000 CLIPPING WINDOW	0010
0101	0100	0110

Fig. 1. Region codes for nine regions

Regions are set up in reference to the boundaries as shown in Figure-1. Each bit position in the region code is used to indicate one of the four relative co-ordinate positions of the point with respect to the clip window: to the left, right, top, or bottom. By numbering the bit positions in the region code as 1 through 4 from right to left, the coordinate regions can be correlated with the bit positions as bit 1 for left, bit 2 for right, bit 3 for below and bit 4 for above. A value of 1 in any bit position indicates that the point is in that relative position; otherwise, the bit position is set to 0. If a point is within the clipping rectangle, the region code is 0000.

3 Intersection Tests

In this section four procedures are defined to test a line for intersection at left, right, bottom and top clipping window edges. Each procedure takes a line as input and returns true if it intersects left, right, bottom or top clipping window edge depending on the procedure used; otherwise it returns false.

The equation of the line with end points (x_1, y_1) and (x_2, y_2) and slope $m = (y_2 - y_1) / (x_2 - x_1)$ is given by: $y - y_1 = m(x - x_1)$

A rectangular clip window is taken with coordinates of the lower left, lower right, upper left and upper right corners as (X_{wmin}, Y_{wmin}) , (X_{wmax}, Y_{wmin}) , (X_{wmin}, Y_{wmax}) and (X_{wmax}, Y_{wmax}) respectively shown in the figure-2.

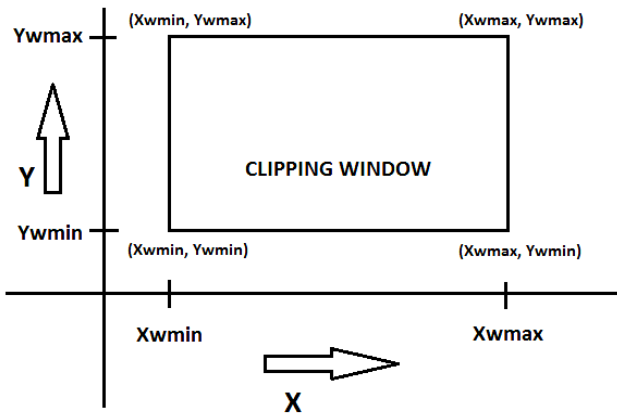


Fig. 2. Coordinates of corners of the clipping window

3.1 Intersection Test at Left Edge

Following procedure tests intersection of a given line at left edge of the clipping window:

```
Procedure IntersectionLeft (Line)
begin
```

```
     $y = y_1 + m (X_{wmin} - x_1);$ 
    if value of  $y$  is inside the range  $(Y_{wmin}, Y_{wmax})$ 
    begin
        the given line intersects the left
        clipping window edge;
        return true;
    end;
    else
    return false;
```

```
end.
```

3.2 Intersection Test at Right Edge

Following procedure tests intersection of a given line at right edge of the clipping window:

```

Procedure IntersectionRight (Line)
begin
     $Y = Y_1 + m (X_{wmax} - x_1)$ ;
    if value of y is inside the range (Ywmin, Ywmax)
    begin
        the given line intersects the right
        clipping window edge;
        return true;
    end;
    else
        return false;
end.

```

3.3 Intersection Test at Bottom Edge

Following procedure tests intersection of a given line at bottom edge of the clipping window:

```

Procedure IntersectionBottom (Line)
begin
     $x = x_1 + (Y_{wmin} - y_1) / m$ ;
    if value of x is inside the range (Xwmin, Xwmax)
    begin
        the given line intersects the bottom clipping
        window edge;
        return true;
    end;
    else
        return false;
end.

```

3.4 Intersection Test at Top Edge

Following procedure tests intersection of a given line at top edge of the clipping window:

```

Procedure IntersectionTop (Line)
begin
     $x = x_1 + (Y_{wmax} - y_1) / m$ ;
    if value of x is inside the range (Xwmin, Xwmax)
    begin
        the given line intersects the top clipping
        window edge;
        return true;
    end;
end.

```

```

end;
else
return false;
end.

```

4 Modified Trivial Rejection Criteria

According to Cohen-Sutherland line clipping algorithm we have total nine regions, out of these nine regions, one region is inside the clipping window with region code 0000. If a line is completely outside the clipping window, no end point of the line belongs to region code 0000. So we are left with only eight regions. We are not entertaining the cases where both end points of a line fall in the same region because we can trivially reject these lines by observing the same code for both end points. Therefore the work focuses on the lines with end points fall in different regions excluding the region with region code 0000. Now with eight regions and two end points of a line we have total ${}^8C_2 = 28$ possible combinations. Here the case of the line with end points (x_1, y_1) and (x_2, y_2) is considered same as the line with end points (x_2, y_2) and (x_1, y_1) . Out of these 28 cases the lines of the two cases, with region codes 0001 & 0010 and 1000 & 0100 can not be completely outside as shown in the figure-3.

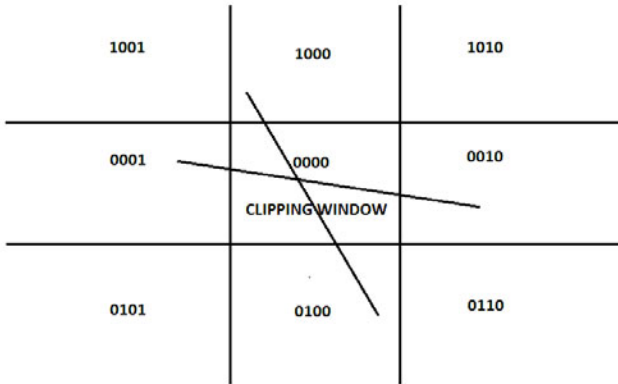


Fig. 3. Exclusions for trivial rejection

Now we are left with 26 cases. Out of these 26 cases, 12 cases have non-zero bitwise AND of both end points and therefore effectively solved by Cohen-Sutherland line clipping algorithm by trivially rejecting the line. For the remaining 14 cases, the bitwise AND of both end points is zero. These 14 cases can be solved by Cohen-Sutherland line clipping algorithm by dividing the line into segments and then examine each segment for acceptance and rejection. This approach consumes more time. Following list in table-1 shows all these 14 cases which will be solved quickly without any segmentation.

Table 1. 14 Cases to be solved without segmenting the line

Case	Line id.	End Point1	End point2	Bitwise OR
1.	B1	0001	0100	0101
2.	B2	0010	0100	0110
3.	T1	0001	1000	1001
4.	T2	0010	1000	1010
5.	LR1	1001	0010	1011
6.	LR2	0001	1010	1011
7.	LR3	0001	0110	0111
8.	LR4	0101	0010	0111
9.	BT1	1001	0100	1101
10.	BT2	1000	0101	1101
11.	BT3	1000	0110	1110
12.	BT4	1010	0100	1110
13.	LRBT1	1001	0110	1111
14.	LRBT2	0101	1010	1111

These 14 cases can be merged in five types. B-type, T-type, LR-type, BT-type and LRBT-type of lines. Now we'll examine the behavior of all the lines fall in these five types separately and provide the solution.

4.1 B-type Lines

B-type consists of two cases, line B1 and line B2 as shown in figure-4.

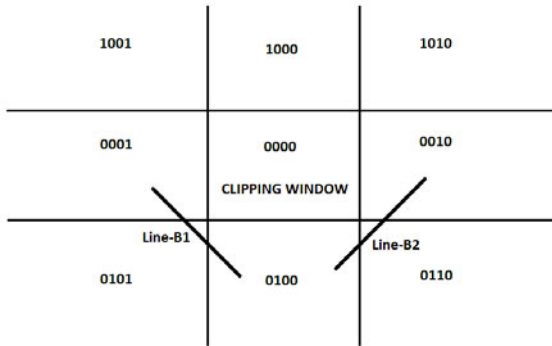


Fig. 4. B-type Lines

We can easily understand the behavior of B-type lines by analyzing figure-4. Both of the B-type lines, i.e., Line B1 and Line B2, if not completely outside the clipping window, then must intersect either left and bottom edge or right and bottom edge of the clipping window. It means both the lines if not completely outside the clipping window must intersect bottom edge of the clipping window. It implies that if B-type lines are outside the clipping window then these lines must not intersect the bottom edge of the clipping window. Therefore B-type lines can be clipped by following tests:

If (bitwise OR of two end points is 0101 or 0110) and (IntersectionBottom (Line) = False), then the given line is trivially rejected. (Passing parameter Line may contain any line of B-type.)

4.2 T-type Lines

T-type also consists of two cases, line T1 and line T2 as shown in figure-5. The case of T-type lines is similar to B-type lines. We can easily understand the behavior of T-type lines by analyzing figure-5. Both of the T-type lines, i.e., Line T1 and Line T2, if not completely outside the clipping window, then must intersect either left and

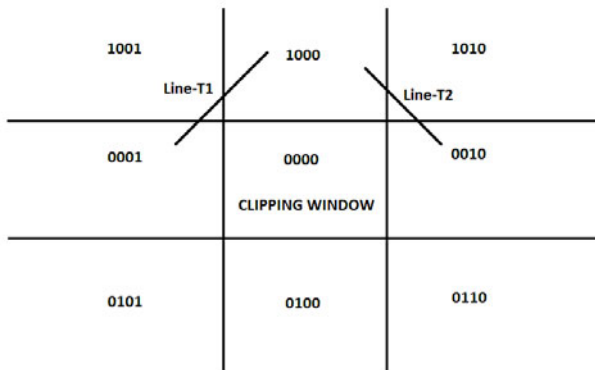


Fig. 5. T-type Lines

top edge or right and top edge of the clipping window. It means both the lines if not completely outside the clipping window must intersect top edge of the clipping window. It implies that if T-type lines are outside the clipping window then these lines must not intersect the top edge of the clipping window.

Therefore T-type lines can be clipped by following tests:

If (bitwise OR of two end points is 1001 or 1010) and (IntersectionTop (Line) = False), then the given line is trivially rejected. (Passing parameter Line may contain any line of T-type.)

4.3 LR-type Lines

LR-type consists of four cases, line LR1, line LR2, Line LR3 and line LR4 as shown in figure-6. We can verify by analyzing figure-6 that if line LR1 is not completely outside the clipping window then it must intersect either left and right clipping window edge or right and top clipping window edge. If line LR2 is not completely outside the clipping window then it must intersect either left and right clipping window edge or left and top clipping window edge. If line LR3 is not completely outside the clipping window then it must intersect either left and right clipping window edge or left and bottom clipping window edge. Similarly if line LR4 is not completely outside the clipping window then it must intersect either left and right clipping window edge or right and bottom clipping window edge. Therefore on summarizing above the conclusion is that if the lines of LR-type do not intersect left and right clipping window edges then the lines are trivially rejected.

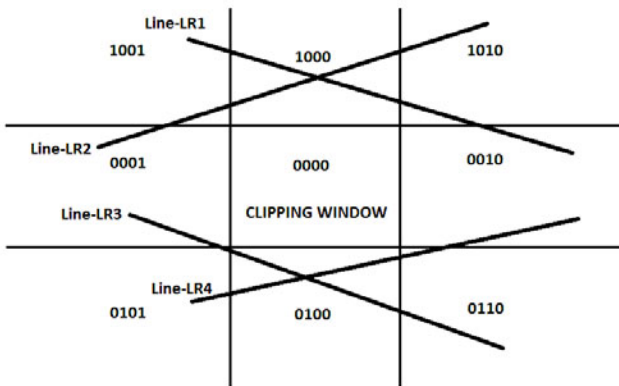


Fig. 6. LR-type Lines

The LR-type lines can be clipped by following tests, OR is the Boolean operator used in the following test.

If (bitwise OR of two end points is 1011 or 0111) and (IntersectionLeft (Line) OR IntersectionRight (Line) = False), then the given line is trivially rejected. (Passing parameter Line may contain any line of LR-type.)

4.4 BT-type Lines

BT-type consists of four cases, line BT1, line BT2, Line BT3 and line BT4 as shown in figure-7. We can verify by analyzing figure-7 that if line BT1 is not completely outside the clipping window then it must intersect either bottom and top clipping window edge or left and bottom clipping window edge. If line BT2 is not completely outside the clipping window then it must intersect either bottom and top clipping window edge or left and top clipping window edge. If line BT3 is not completely outside the clipping window then it must intersect either bottom and top clipping window edge or right and top clipping window edge. Similarly if line BT4 is not completely outside the clipping window then it must intersect either bottom and top clipping window edge or right and bottom clipping window edge. Therefore on summarizing above the conclusion is that if the lines of BT-type do not intersect bottom and top clipping window edges then the lines are trivially rejected.

The BT-type lines can be clipped by following tests, OR is the Boolean operator used in the following test.

If (bitwise OR of two end points is 1101 or 1110) and (IntersectionBottom (Line) OR IntersectionTop (Line) = False), then the given line is trivially rejected. (Passing parameter Line may contain any line of BT-type.)

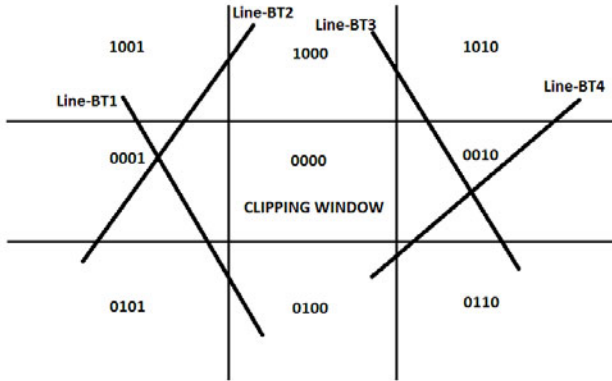


Fig. 7. BT-type Lines

4.5 LRBT-type Lines

LRBT-type consists of two cases, line LRBT1 and line LRBT2 as shown in figure-8.

The last LRBT-type is the most complex one because the lines of this type if not completely outside the clipping window then these lines may intersect the clipping window edges in four different ways. The lines of case LRBT-1 if not completely outside the clipping window can intersect either left and bottom edge or left and right edge or top and bottom edge or top and right edge of the clipping window. Similarly the lines of the case LRBT-2 if not completely outside the clipping window can intersect either left and top edge or left and right edge or top and bottom edge or bottom and right edge of the clipping window.

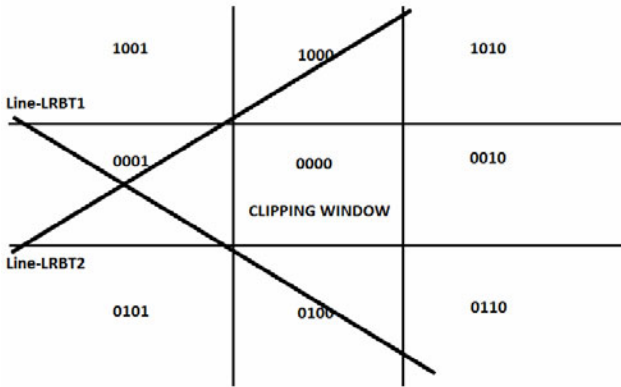


Fig. 8. LRBT-type Lines

Therefore the test for these lines is as follows, OR is the Boolean operator used in the following test.

If (bitwise OR of two end points is 1111) and (IntersectionLeft (Line) OR IntersectionRight (Line) OR IntersectionBottom (Line) OR IntersectionTop (Line) = False), then the given line is trivially rejected. (Passing parameter Line may contain any line of LRBT-type.)

5 Conclusion

A Novel criterion for trivial rejection (total clipping) of a line has been proposed in this paper. The proposed criterion doesn't require dividing the line into segments. Total 14 cases were explored for total clipping and tests were provided for all the cases. We claim that all the lines of these 14 cases can be totally clipped by the provided mechanism.

References

1. Hearn, D., Baker, M.P.: Computer Graphics, 2nd edn., pp. 226–228 (1994)
2. Hearn, D., Baker, M.P.: Computer Graphics C Version, 2nd edn., pp. 225–228 (1997)
3. Foley, J.D., Dam, A., Feiner, S.K., Hughes, J.F.: Computer Graphics Principles and Practice, 2nd edn., pp. 111–115 (1996)
4. Rogers, D.F.: Procedural Elements for Computer Graphics, 2nd edn., pp. 181–183 (1998)
5. Harrington, S.: Computer Graphics A Programming Approach, 2nd edn., pp. 181–183 (1987)
6. Devai, F.: An Analysis Technique and an Algorithm for Line Clipping
7. Bhuiyan, M.M.I.: Designing a Line Clipping Algorithm by Categorizing Line Dynamically and Using Intersection Point Method. In: International Conference on Electronic Computer Technology (2009)
8. Huang, W., Wangyong: A Novel Algorithm for Line Clipping

Greedy Views Selection Using Size and Query Frequency

T.V. Vijay Kumar and Mohammad Haider

School of Computer and Systems Sciences,
Jawaharlal Nehru University,
New Delhi-110067, India

Abstract. Greedy view selection, in each iteration, selects the most beneficial view for materialization. Algorithm HRUA, the most fundamental greedy based algorithm, uses the size of the views to select the top-k beneficial views from a multidimensional lattice. HRUA does not take into account the query frequency of each view and as a consequence it may select views which may not be beneficial in respect of answering future queries. As a result, the selected views may not contain relevant and required information for answering queries leading to an unnecessary space overhead. This problem is addressed by the algorithm proposed in this paper, which considers both the size and the query frequency of each view to select the top-k views. The views so selected are profitable with respect to size and are capable of answering large number of queries. Further, experiments show that the views selected using the proposed algorithm, in comparison to those selected using HRUA, are able to answer comparatively greater number of queries at the cost of a slight drop in the total cost of evaluating all the views. This in turn aids in reducing the query response time and facilitates decision making.

Keywords: Materialized Views, View Selection, Greedy Algorithm.

1 Introduction

Historical data has been used by industries to evolve business strategies in order to be competitive in the market. Data warehouse stores such historical data on which analytical queries are posed for strategic decision making [7]. The size of the data warehouse, which continuously grows with time, and the nature of analytical queries, which are long and complex, leads to high query response time. This query response time needs to be reduced in order to make decision making more efficient. One way to address this problem is by answering queries using materialized views, which are pre-computed and summarized information stored in a data warehouse[9]. Their aim is to reduce the response time for analytical queries.

The number of possible views is exponential in the number of dimensions and therefore all cannot be materialized due to limitation in storage space available for view materialization [6]. Thus, there is a need to select a subset of views from among all possible views that improves the query response time. Selecting an optimal subset of such views is shown to be an NP-Complete problem [6]. Further, materialized views cannot be arbitrarily selected as they are required to contain information that is useful for answering future queries resulting in reduced response time. This problem

is referred to as view selection problem in literature [4]. Several view selection algorithms have been proposed in literature, most of which are greedy based [1, 2, 3, 5, 6, 8, 10, 11, 13, 14]. The greedy based view selection, in each iteration, selects the most beneficial view for materialization. Most of the greedy algorithms are focused around the algorithm in [6], which hereafter in this paper will be referred to as HRUA. HRUA selects top-k beneficial views from a multidimensional lattice. It is based on a linear cost model, where the cost is in terms of the size of the view. This cost is used to compute the benefit of each view as given below:

$$\text{Benefit } V = \sum \{ (\text{Size}(\text{SMA}(W)) - \text{Size}(V)) \mid V \text{ is an ancestor of view } W \text{ in the lattice and } (\text{Size}(\text{SMA}(W)) - \text{Size}(V)) > 0 \}$$

where $\text{Size}(V) = \text{Size of view } V$

$\text{Size}(\text{SMA}(V)) = \text{Size of Smallest Materialized Ancestor of view } V$.

Though HRUA uses size of the view to compute its benefit, it does not take into account the query frequency of each view, which specifies the number of queries that can be answered by a view. As a consequence, HRUA may select views that may not be beneficial in respect of answering future queries. This in turn would result in the selected views using space without having relevant and required information for answering queries. As an example, consider a three dimensional lattice shown in Fig. 1(a). The size of the view in million (M) rows, and the query frequency (QF) of each view, is given alongside the view. Selection of Top-3 views using HRUA is shown in Fig. 1(b).

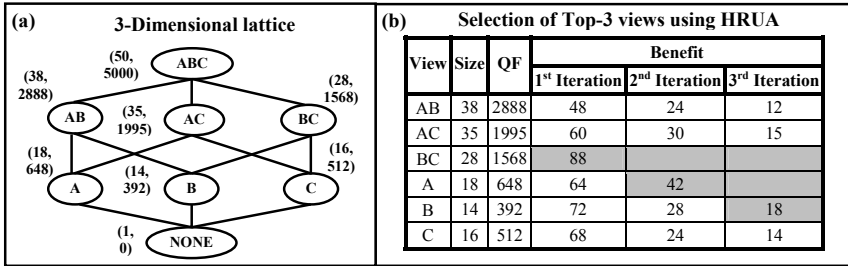


Fig. 1. Selection of Top-3 views using HRUA

HRUA assumes the root view to be materialized as queries on it are unlikely to be answered by any other views in the lattice. HRUA selects A, AB and C as the Top-3 views. These selected views result in a Total View Evaluation Cost (TVEC) of 252. If the query frequency of each view is considered, the Total Queries Answered (TQA) by the selected views is 3120, from among 8003 queries. This TQA value needs to be improved upon so that greater number of queries can be answered. The algorithm presented in this paper attempts to improve this TQA value by considering both the size, and the query frequency of each view to select the Top-k profitable views for materialization. The proposed algorithm aims to select views that are profitable with respect to size and also provide answers to large number of queries.

The paper is organized as follows: The proposed algorithm is given in section 2 followed by examples based on it in section 3. The experimental results are given in section 4. Section 5 is the conclusion.

2 Proposed Algorithm

As mentioned above, HRUA selects views that are beneficial with respect to size but may be unable to answer large number of queries. As a consequence, the query response time may be high. This problem can be addressed if selected views take into account not only their size but also their ability to answer queries, i.e. query frequency. The proposed algorithm aims to select such views by considering query frequency, along with the size, of the view to select the most profitable views for materialization. The proposed algorithm assumes that past queries provide useful indicators of queries likely to be posed in future and thus use them to determine the query frequency of each view. The proposed algorithm, as given in Fig. 2, takes the lattice of views along with the size and query frequency of each view as input and produces the Top-K views as output.

```

Input: lattice of views L along with size and query frequency of each view
Output: Top-k views
Method:
  Let
     $V_R$  be the root view in the lattice,  $S(V)$  be the size of view  $V$ ,  $QF(V)$  be the query frequency of  $V$  in the lattice,
     $SMA(V)$  be the smallest materialized ancestor of  $V$ ,  $D(V)$  be the set of all descendent views of  $V$ ,  $MV$  be the set
    of materialized views,  $P(V) = \text{Profit of view } V$ ,  $P_M = \text{Maximum Profit}$ ,  $V_P = \text{View with maximum profit}$ 
  FOR  $V \in L$ 
     $SMA(V) = \text{RootView}$ 
  END FOR
  REPEAT
     $P_M = 0$ 
    FOR each view  $V \in (L - V_R \cup MV)$ 
       $V_P = V$ 
       $P(V) = 0$ 
      FOR each view  $W \in D(V)$  and  $(S(SMA(W)) - S(V)) > 0$ 
        
$$P(V) = P(V) + \left| \frac{QF(SMA(W))}{S(SMA(W))} - \frac{QF(V)}{S(V)} \right|$$

      END FOR
      IF  $P_M < P(V)$ 
         $P_M = P(V)$ 
         $V_P = V$ 
      END IF
    END FOR
     $MV = MV \cup \{V_P\}$ 
    FOR  $W \in D(V_P)$ 
      IF  $S(SMA(W)) > S(V_P)$ 
         $SMA(W) = V_P$ 
      END IF
    END FOR
  Until  $|MV| < k$ 
  Return  $MV$ 

```

Fig. 2. Proposed Algorithm

The proposed algorithm, in each iteration, computes the profit of each view $P(V)$ as given below:

$$P(V) = \sum \left\{ \left| \frac{QF(SMA(W))}{S(SMA(W))} - \frac{QF(V)}{S(V)} \right| \mid V \text{ is an ancestor of view } W \text{ in the lattice and } (S(SMA(W)) - S(V)) > 0 \right\}$$

The profit of a view V is computed as the product of the number of dependents of V and the query frequency per unit size difference of V with its smallest materialized ancestor. The profit of each, as yet unselected view, in each iteration and select the most profitable view from amongst them for materialization. In this way, the proposed algorithm continues to select top profitable view until K views are selected.

Examples illustrating selection of views using the Proposed Algorithm (PA) are given next.

3 Examples

Let us consider selection of the Top-3 views from the multidimensional lattice in Fig. 1(a) using the proposed algorithm. The selection of Top-3 views is given in Fig. 3.

View	Size	QF	Profit		
			1 st Iteration	2 nd Iteration	3 rd Iteration
AB	38	2888	96	48	24
AC	35	1995	172	86	
BC	28	1568	176		
A	18	648	128	84	41
B	14	392	144	56	56
C	16	512	136	48	48

Fig. 3. Selection of Top-3 views using PA

PA selects AC, A and BC as the Top-3 views. These selected views have a TVEC value of 254 and a TQA value of 5115. Though the TVEC value (254) of views selected using PA is slightly inferior to the TVEC value (252) of views selected using HRUA, the views selected using PA have a significantly higher value of TQA (5115), when compared with the TQA value (3120) of views selected using HRUA. That is, the views selected using PA are able to account for a greater number of queries at the cost of a slight increase in the TVEC value.

PA may also select views that not only account for more number of queries but also may have lesser or better TVEC. As an example, consider a three dimensional lattice shown in Fig. 4(a).

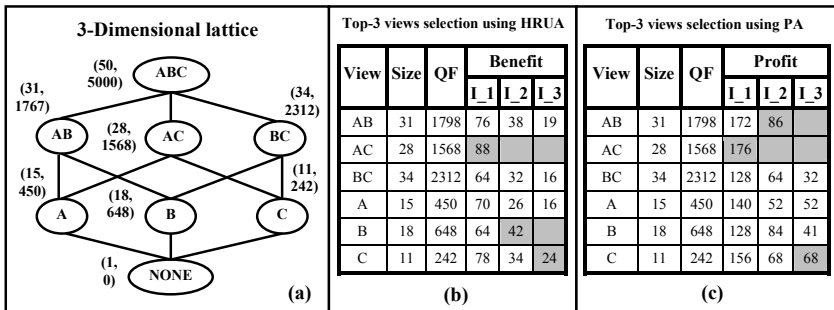


Fig. 4. Selection of Top-3 views using HRUA and PA

HRUA selects AC, B and C as the Top-3 views as against AC, AB and C selected by PA. The views selected using PA has TVEC of 240, which is less than TVEC of

246 due to views selected using HRUA. Also, the views selected using PA has comparatively higher value of TQA of 4675 against the TQA of 2908 due to views selected using HRUA. Thus, it can be said that PA, in comparison to HRUA, is capable of selecting views that not only account for greater number of queries but also at lower total cost of evaluating all the views.

In order to compare the performance of PA with respect to HRUA, both the algorithms were implemented and run on data sets with varying dimensions. The experiment based comparisons of PA and HRUA are given next.

4 Experimental Results

The PA and HRUA algorithms were implemented using JDK 1.6 in Windows-XP environment. The two algorithms were experimentally compared on an Intel based 2 GHz PC having 1 GB RAM. The comparisons were carried out on parameters like TVEC and TQA for selecting Top-20 views for materialization. The experiments were conducted by varying the number of dimensions of the data set from 4 to 10.

First, graphs were plotted to compare PA and HRUA algorithms on TQA against the number of dimensions. The graphs are shown in Fig. 5. It is observed from the graph (Fig. 5(a)) that the increase in TQA, with respect to number of dimensions, is higher for PA vis-à-vis HRUA. This difference even exists for 4 to 7 dimensions as evident in the graph shown in Fig. 5(b).

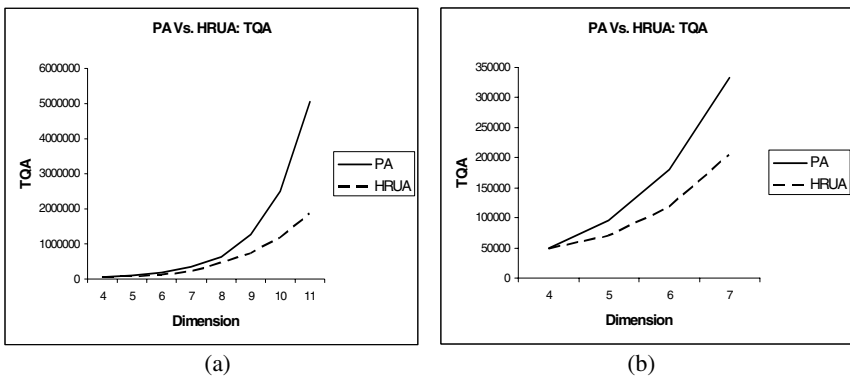


Fig. 5. TQA - PA Vs. HRUA

In order to ascertain the impact of better TQA, due to PA, on the TVEC, graphs for TQA against number of dimensions were plotted and are shown in Fig. 6. It is evident from the graph (Fig. 6(a)) that the TVEC of PA is slightly more than that of HRUA. This difference is almost negligible for dimensions 4 to 7 as shown in Fig. 6(b). This small difference shows that the PA selects views which are almost similar in quality to those selected by HRUA.

It can be reasonably inferred from the above graphs that PA trades significant improvement in TQA with a slight drop in TVEC of views selected for materialization.

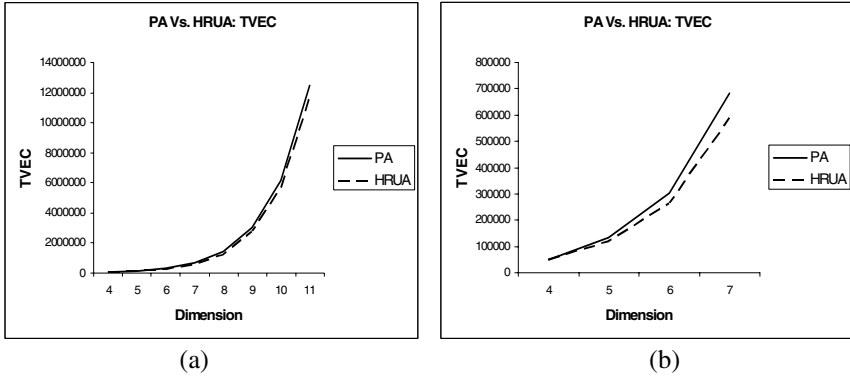


Fig. 6. TVEC - PA Vs. HRUA

5 Conclusion

In this paper, an algorithm is proposed that greedily selects Top-K views from a multidimensional lattice using views size and query frequency. The algorithm computes the profit of each view, which is defined as a function of the size and query frequency, and then selects from amongst them the most profitable view for materialization. Unlike HRUA, the proposed algorithm is able to select views that are not only profitable with respect to size but are also able to account for large number of queries. The selected views thereby would reduce the average query response time.

Further experiment based comparison between the proposed algorithm and HRUA on parameters TQA and TVEC showed that the proposed algorithm, in comparison to HRUA, was found to achieve significant improvement in TQA at the cost of a slight drop in the TVEC in respect of views selected for materialization. This shows that the proposed algorithm trades significant improvement in total number of queries answered with a slight drop in the quality of views selected for materialization.

References

1. Agrawal, S., Chaudhuri, S., Narasayya, V.: Automated Selection of Materialized Views and Indexes in SQL Databases. In: Proceedings of VLDB 2000, pp. 496–505. Morgan Kaufmann Publishers, San Francisco (2000)
2. Aouiche, K., Darmoni, 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: Proceedings of VLDB 1997, pp. 156–165. Morgan Kaufmann Publishers, San Francisco (1997)
4. Chirkova, R., Halevy, A., Suciu, D.: A Formal Perspective on the View Selection Problem. *The VLDB Journal* 11(3), 216–237 (2002)
5. Gupta, H., Mumick, I.: Selection of Views to Materialize in a Data Warehouse. *IEEE Transactions on Knowledge and Data Engineering* 17(1), 24–43 (2005)

6. Harinarayan, V., Rajaraman, A., Ullman, J.: Implementing Data Cubes Efficiently. In: Proceedings of SIGMOD 1996, pp. 205–216. ACM Press, New York (1996)
7. Inmon, W.H.: Building the Data Warehouse, 3rd edn. Wiley Dreamtech (2003)
8. Nadeau, T.P., Teorey, T.J.: Achieving scalability in OLAP materialized view selection. In: Proceedings of DOLAP 2002, pp. 28–34. ACM, New York (2002)
9. Roussopoulos, N.: Materialized Views and Data Warehouse. In: 4th Workshop KRDB 1997, Athens, Greece (August 1997)
10. Serna-Encinas, M.T., Hoya-Montano, J.A.: Algorithm for selection of materialized views: based on a costs model. In: Proceeding of Eighth International Conference on Current Trends in Computer Science, pp. 18–24 (2007)
11. Shah, A., Ramachandran, K., Raghavan, V.: A Hybrid Approach for Data Warehouse View Selection. *Int. Journal of Data Warehousing and Mining* 2(2), 1–37 (2006)
12. Teschke, M., Ulbrich, A.: Using Materialized Views to Speed Up Data Warehousing. Technical Report, IMMD 6, Universität Erlangen-Nürnberg (1997)
13. Vijay Kumar, T.V., Ghoshal, A.: A Reduced Lattice Greedy Algorithm for Selecting Materialized Views. In: CCIS, vol. 31, pp. 6–18. Springer, Heidelberg
14. Vijay Kumar, T.V., Haider, M., Kumar, S.: Proposing Candidate Views for Materialization. In: CCIS, vol. 54, pp. 89–98. Springer, Heidelberg (2010)

Quantum-Inspired Differential Evolution on Bloch Coordinates of Qubits

Ankit Pat¹, Ashish Ranjan Hota², and Avneet Singh²

¹ Department of Mathematics

² Department of Electrical Engineering

Indian Institute of Technology, Kharagpur, India

{ankitpat.iitkgp, avneet.singh.16}@gmail.com,

hota.ashish@acm.org

Abstract. Differential evolution (DE) is a population based evolutionary algorithm widely used for solving multidimensional global optimization problems over continuous spaces. On the other hand, a variation of the original quantum-inspired evolutionary algorithm (QEA), bloch quantum-inspired evolutionary algorithm (BQEA), is a promising concept which very well suitable for handling global optimization problem of low dimensionality. BQEA applies several quantum computing techniques such as qubit representation based on bloch sphere and rotation gate operator, etc. This paper extends the concept of differential operators to the quantum paradigm and proposes the bloch quantum-inspired differential evolution algorithm (BQDE). The performance of BQDE is found to be significantly superior as compared to BQEA on several benchmark functions.

Keywords: Differential Evolution, Bloch coordinates, Quantum-inspired Evolutionary Algorithms, Quantum Computing.

1 Introduction

Differential Evolution (DE) introduced by Storn and Price [1,2] has been shown to give significantly better performance in terms of efficiency and robustness on several benchmark multimodal continuous functions than other population based evolutionary algorithms. A large number of modifications have been proposed to make the selection of control parameters of DE adaptive and free from function dependency [3-6].

To solve various optimization problems better than the conventional evolutionary algorithms, a broad class of algorithms have been proposed by applying several concepts of quantum computing in the past decade. Thus quantum inspired genetic algorithms with interference as crossover operator [7], quantum inspired evolutionary algorithms (QEA) [8], quantum behaved particle swarm optimization [9] etc has been developed for both continuous and binary spaces.

QEAs have been extended by differential operators to solve flow shop scheduling problems [10], N-queen's problem [11], for classification rule discovery [12] and

some benchmark functions [13]. All of these use the qubit representation scheme and as shown in [8], it is used to solve continuous optimization problems by discretizing the solution space. Recently, a novel representation scheme based on Bloch Sphere was proposed [14], which represented the solution in continuous space and thus suitable for functional optimization. In this paper, a new bloch quantum-inspired differential evolution algorithm (BQDE) is proposed with a novel mutation scheme. The proposed BQDE outperforms BQEA and DE over a wide variety of benchmark functions.

The rest of this paper is organized as follows: Section 2 gives a brief introduction of continuous optimization problem, DE and BQEA. The proposed BQDE is explained in detail in section 3. Experimental settings and the results obtained are mentioned under section 4. Finally, section 5 concludes the paper.

2 Background

2.1 Differential Evolution

In classical DE, each member of the population is represented by a real valued D-dimensional vector. One iteration of the DE algorithm consists of three major operations – mutation, crossover and selection, which are carried out for each member of the population (called as target vector). The three operations are discussed in detail below.

Mutation: The mutant V_i^t vector on a target vector X_i^t is generated by adding a randomly selected vector X_{r1}^t from the population, with a weighted difference of two other randomly selected vectors X_{r2}^t, X_{r3}^t from the population. The selected vectors must be distinct among themselves and also from the target vector.

$$V_i^t = X_{r1}^t + F.(X_{r2}^t - X_{r3}^t) \quad (1)$$

Crossover: The crossover operation generates a trial vector U_i from its corresponding target vector X_i and mutant vector V_i , by using the following relation:

$$u_{j,i}^t = \begin{cases} v_{j,i}^t, & \text{if } (rand_j(0,1) \leq CR) \text{ or } (j = I_{rand}) \\ x_{j,i}^t, & \text{if } (rand_j(0,1) > CR) \text{ and } (j \neq I_{rand}) \end{cases} \quad (2)$$

where $j=1,2,\dots,D$, $U_i = (u_{1,i}^t, u_{2,i}^t, \dots, u_{D,i}^t)$, $rand_j$ is the j th evaluation of a random number generator in $[0,1]$ from a uniform distribution. I_{rand} is a randomly chosen dimension index from $\{1,2,\dots,D\}$ which ensures that the new trail vector is different from the target vector. CR is a control parameter which decides the crossover rate and its value is typically chosen between 0 and 1.

Selection: If the trial vector U_i has a better fitness value compared to the target vector, then it replaces the target vector in the population in the next iteration. Otherwise, the target vector remains unchanged in the population.

2.2 Bloch Quantum-Inspired Evolutionary Algorithm

This sub-section describes the Bloch Quantum-inspired Evolutionary Algorithm.

Representation: Instead of using $[\alpha, \beta]^T$ like QEA as the representation of Q-bits, BQEA uses the bloch sphere based representation as follows:

$$|q\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\varphi} \sin\left(\frac{\theta}{2}\right)|1\rangle \quad (3)$$

$$\text{where } \left|\cos\left(\frac{\theta}{2}\right)\right|^2 + \left|e^{i\varphi} \sin\left(\frac{\theta}{2}\right)\right|^2 = 1 \quad (4)$$

The real φ and θ correspond to a point P on the Bloch sphere. Thus a qubit can be described by bloch coordinates as $[\cos \varphi \sin \theta, \sin \varphi \sin \theta, \cos \theta]^T$. In BQEA, the qubits are directly considered as genes and the chromosome looks like as below:

$$p_i = \begin{bmatrix} \cos \varphi_{i1} \sin \theta_{i1} & \cos \varphi_{i2} \sin \theta_{i2} & \dots & \cos \varphi_{iD} \sin \theta_{iD} \\ \sin \varphi_{i1} \sin \theta_{i1} & \sin \varphi_{i2} \sin \theta_{i2} & \dots & \sin \varphi_{iD} \sin \theta_{iD} \\ \cos \theta_{i1} & \cos \theta_{i2} & \dots & \cos \theta_{iD} \end{bmatrix} \quad (5)$$

where $\varphi_{ij} = 2\pi * \text{rnd}$ and $\theta_{ij} = \pi * \text{rnd}$, rnd is a random number uniformly distributed between 0 and 1. $i=1, 2, \dots, n$ and $j=1, 2, \dots, D$, where n is the number of chromosomes or the population size and D is the number of dimensions. The Bloch coordinates of each qubit are regarded as three paratactic genes, each chromosome contains three paratactic gene chains, and each gene chain represents an optimization solution. Thus each chromosome represents three optimization solutions simultaneously. This makes the search process faster and also avoids multiple decoding of the probability amplitude to binary values as in QEA [8].

Solution space Transform: A linear transform is used to transform the gene chains to the solution space. If the jth qubit of chromosome pi is $[x_{ij}, y_{ij}, z_{ij}]^T$, then the corresponding variables in the solution space are as given below:

$$X_{i\alpha}^j = \frac{b_j(1 + \alpha_{ij}) + a_j(1 - \alpha_{ij})}{2}, \quad (6)$$

where $i=1, 2, \dots, n$, $\alpha=x, y$ and z . $j=1, 2, \dots, D$. a_j and b_j are the upper and lower limits of the solution space.

Update: The chromosomes are updated using a rotation gate operator as defined in [14]. The values of $\Delta\theta$ and $\Delta\phi$ are determined by comparing the relative phase between the present solution and the current optimum solution using a query table [8].

Mutation: To avoid premature convergence and to escape from local optimum solutions, quantum non-gate is used to introduce large changes in phase magnitudes of the qubits. The operation uses the mutation operator as defined in [14]. Mutation operator is applied on an individual qubit with a certain probability known as mutation probability p_m .

3 Bloch Quantum-Inspired Differential Evolution

In the proposed Bloch Quantum Differential Evolution (BQDE) algorithm, the representation of qubits and transforming them to the solution space are identical to that of BQEA. But, the update and mutation operators are replaced by a novel differential mutation and crossover. This is followed by a greedy selection strategy. The detailed descriptions of the last three operations are as follows.

Mutation: Mutation operator applies to each chromosome in the population and generates a mutant chromosome. For this operation, three chromosomes are randomly selected from the population, which are mutually distinct and also different from the original chromosome. Then, the corresponding φ and θ of the mutant chromosome are generated as follows:

$$\theta_{ij}^m = \frac{\theta_{r_1j} + F1 \cdot \theta_{r_2j} + F2 \cdot \theta_{r_3j}}{1 + F1 + F2} \quad (7)$$

$$\varphi_{ij}^m = \frac{\varphi_{r_1j} + F1 \cdot \varphi_{r_2j} + F2 \cdot \varphi_{r_3j}}{1 + F1 + F2} \quad (8)$$

where F1 and F2 are two random numbers uniformly distributed between -1 and 1 and F1+F2 is not equal to -1.

Crossover: The crossover operation operates on the original qubits and the respective mutant qubits in the following manner:

$$\theta_{ij}^c = \begin{cases} \theta_{ij}^m, & \text{if } (rand_j(0,1) \leq CR^t) \text{ or } (j = I_{rand}) \\ \theta_{ij}, & \text{if } (rand_j(0,1) > CR^t) \text{ and } (j \neq I_{rand}) \end{cases} \quad (9)$$

$$\varphi_{ij}^c = \begin{cases} \varphi_{ij}^m, & \text{if } (rand_j(0,1) \leq CR^t) \text{ or } (j = I_{rand}) \\ \varphi_{ij}, & \text{if } (rand_j(0,1) > CR^t) \text{ and } (j \neq I_{rand}) \end{cases} \quad (10)$$

where θ_{ij}^c and φ_{ij}^c are the jth qubit of ith individuals after the crossover operation. I_{rand} is a number randomly chosen from $\{1,2,\dots,D\}$ which ensures at least one qubit is different from the original set in each individual. CR^t is the control parameter which is determined in every iteration as follows:

$$CR^t = 0.07 + 0.04 \cdot rand_t(0,1) \quad (11)$$

This value was selected because it was found to give better solution experimentally in only in this small interval.

Selection: The qubits obtained after crossover are evaluated by solution space transformation. If the fitness value of any of the three solutions corresponding to the crossover qubit is higher than the solutions corresponding to the original qubit, then the original qubit is replaced by the new set of qubits.

The pseudo-code of BQDE is provided below.

Procedure BQDE

begin

$t \leftarrow 0$

initialize $Q(t)$ by (5)

make $X(t)$ from $Q(t)$ by (6)

evaluate $X(t)$ by function evaluation

$BX \leftarrow$ best solution among $X(t)$

$BC \leftarrow$ chromosome corresponding to BX

$GX \leftarrow BX$

$GC \leftarrow BC$

while $t < T$ **do**

$t \leftarrow t+1$

generate mutant qubits by (7,8)

evaluate CR^t by (11)

do crossover by (9,10)

select the new qubit as per sec. 3.3

make $X(t)$ from $Q(t)$ by (6)

evaluate $X(t)$

update BX and BC accordingly

if $\text{fit}(BX) < \text{fit}(GX)$

$BX \leftarrow GX$; $BC \leftarrow GC$

else

$GX \leftarrow BX$; $GC \leftarrow BC$

endif

end while

end

4 Experimental Settings and Results

To test the performance of BQDE, five benchmark functions are used (Table 1). All these functions are minimization problems with minimum value at zero. The results of all these functions are compared with basic BQEA, and BQDE (Table 2,3). Asymmetric initialization ranges were chosen for all experiments on all functions [3,4]. The fitness function used is the corresponding functional value. 30 trials were conducted for all experiments and the mean best functional values and respective standard deviation were recorded. Results for all functions have been obtained for populations sizes 20, 40 and 80. For all these cases, maximum number of generations was kept at 2000 and dimension sizes 2 and 5 were considered. The tables 2 and 3 show the performance comparison of the functions as defined in table 1. The values in the table show the mean of the best solution over 30 trial runs and the corresponding standard deviation is shown in the bracket.

From the results, it can be observed that BQDE outperforms BQEA significantly for Sphere, Rastrigin and Ackley's function in all the experimental cases considered. For Rosenbrock function, BQDE outperforms BQEA for dimension two, but in

Table 1. Numerical Benchmark Functions for Performance Evaluation

Function Name	Formulation	Initialization Range	Optimum	Max. Range
Sphere Function	$\sum_{i=1}^n x_i^2$	[50,100]	0	100
Rosenbrock Function	$\sum_{i=1}^n (100.(x_{i+1} - x_i^2)^2 + (x_i - 1)^2)$	[15,30]	0	100
Rastrigin Function	$\sum_{i=1}^n (x_i^2 - 10\cos(2\pi x_i) + 10)$	[2.56,5.12]	0	10
Griewank Function	$\frac{1}{4000} \sum_{i=1}^n x_i^2 - \prod_{i=1}^n \cos(\frac{x_i}{\sqrt{i}}) + 1$	[300,600]	0	600
Ackley's Function	$-20 \exp\left(-0.2 \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}\right) - \exp\left(\frac{1}{n} \sum_{i=1}^n \cos 2\pi x_i\right) + 20 + e$	[16,32]	0	32

Table 2. Performance Evaluation for Sphere and Rosebrock Functions

Pop	Dim	Sphere Function		Rosenbrock Function	
		BQEA	BQDE	BQEA	BQDE
20	2	0.0118 (0.0328)	0.0099 (0.0168)	0.2023 (0.2617)	0.1351 (0.1573)
	5	1.9603 (2.0254)	0.2462 (0.2694)	11.5043 (12.433)	140.70 (332.43)
40	2	0.0190 (0.0389)	0.0030 (0.0033)	0.1001 (0.0896)	0.0195 (0.0210)
	5	1.0240 (1.3280)	0.1242 (0.1288)	5.2803 (3.4813)	23.1715 (14.2110)
80	2	0.0380 (0.0648)	0.0016 (0.0016)	0.0524 (0.0682)	0.0403 (0.0425)
	5	0.1581 (0.1195)	0.1401 (0.1205)	3.6440 (3.4148)	15.1013 (20.6190)

Table 3. Performance Evaluation for Rastrigin, Griewank and Ackley Functions

Pop	Dim	Rastrigin Function		Griewank Function		Ackley Function	
		BQEA	BQDE	BQEA	BQDE	BQEA	BQDE
20	2	0.3795 (0.3529)	0.0021 (0.0022)	0.0425 (0.0457)	0.0195 (0.0220)	0.3284 (0.5510)	0.0921 (0.1393)
	5	7.1080 (3.6924)	0.0466 (0.0417)	0.2895 (0.1539)	0.2032 (0.0987)	1.6733 (0.7770)	0.3226 (0.1413)
40	2	0.2618 (0.3521)	0.0026 (0.0036)	0.0304 (0.0475)	0.0245 (0.0166)	0.0569 (0.1503)	0.0799 (0.0774)
	5	5.7524 (2.1265)	0.1043 (0.1897)	0.2134 (0.0842)	0.1758 (0.0773)	0.8862 (0.6608)	0.6378 (0.4584)
80	2	0.0248 (0.0313)	0.0021 (0.0022)	0.0118 (0.0107)	0.0154 (0.0088)	0.1272 (0.1380)	0.0409 (0.0390)
	5	2.8212 (0.8629)	0.1308 (0.0899)	0.1923 (0.0870)	0.1926 (0.0731)	0.7422 (0.7310)	0.4056 (0.2259)

dimension five , performance of BQDE is very poor and it got trapped in a local minima in most of the runs. For Griewank function, both BQEA and BQDE performed equally well and results were comparable. However, it must be noted that the performance of the both the algorithms degrades as the dimensionality of the problem increases.

5 Conclusion

In this paper, we have proposed a novel BQDE algorithm for solving continuous optimization problems. The proposed algorithm is a hybrid of BQEA and DE along with a novel mutation scheme. The experimental results have proved the superior performance of BQDE compared to BQEA on several benchmark functions. Our future work shall mostly be directed towards designing a self-adaptive method which would generate better solutions across a wide spectrum of benchmark functions.

References

1. Price, K., Storn, R.: Differential Evolution – a simple and efficient adaptive scheme for global optimization over continuous spaces. Technical Report, International Computer Science Institute, Berkeley (1995)
2. Storn, R., Price, K.: Differential Evolution – a simple and efficient Heuristic for global optimization over continuous spaces. *Journal Global Optimization* 11, 341–359 (1997)
3. Teo, J.: Exploring Dynamic Self-adaptive Populations in Differential Evolution. *Soft Computing - A Fusion of Foundations, Methodologies and Applications* 10(8), 673–686 (2006)
4. Brest, J., Greiner, S., Bošković, B., Mernik, M., Žumer, V.: Self- Adapting Control Parameters in Differential Evolution: A Comparative Study on Numerical Benchmark Problems. *IEEE Transactions on Evolutionary Computation* 10(6), 646–657 (2006)
5. Yang, Z., Tang, K., Yao, X.: Self-adaptive Differential Evolution with Neighborhood Search. In: *Proc. IEEE Congress on Evolutionary Computation*, Hong Kong, pp. 1110–1116 (2008)
6. Chakraborty, U.K.: *Advances in Differential Evolution*. Springer, Heidelberg (2008)
7. Narayanan, A., Moore, M.: Quantum-inspired genetic algorithms. In: *Proc. 1996 IEEE Int. Conf. Evolutionary Computation*, Piscataway, NJ, pp. 61–66 (1996)
8. Han, K.-H., Kim, J.-H.: Quantum-inspired evolutionary algorithm for a class of combinatorial optimization. *IEEE Transaction on Evolutionary Computation* 6(6), 580–593 (2002)
9. Sun, J., Feng, B., Xu, W.B.: Particle swarm optimization with particles having quantum behavior. In: *Proc. of the IEEE Congress on Evolutionary Computation*, pp. 325–331 (2004)
10. Jiao, B., Gu, X., Xu, G.: An Improved Quantum Differential Algorithm for Stochastic Flow Shop Scheduling Problem. In: *Proc. IEEE International Conference on Control and Automation*, pp. 1235–1240 (2009)
11. Draa, A., Meshoul, S., Talbi, H., Batouche, M.: A Quantum-Inspired Differential Evolution Algorithm for Solving the N-Queens Problem. *The International Arab Journal of Information Technology* 7(1), 21–27 (2010)
12. Su, H., Yang, Y., Zhao, L.: Classification rule discovery with DE/QDE algorithm. *Expert Systems with Applications* 37, 1216–1222 (2010)
13. Su, H., Yang, Y.: Quantum-inspired differential evolution for binary optimization. In: *The 4-th International Conference on Natural Computation*, pp. 341–346 (2008)
14. Li, P., Li, S.: Quantum-inspired evolutionary algorithm for continuous space optimization based on Bloch coordinates of qubits. *Neurocomputing* 72(1-3), 581–591 (2008)

Extending Speech-Act Based Communication to Enable Argumentation in Cognitive Agents

Punam Bedi and Pooja Vashisth

Department of Computer Science, University of Delhi, Delhi, India
91-011-27667591, 91-011-25990174
punambedi@gmail.com, poojavashisth@rediffmail.com

Abstract. Now days, there is an increasing level of interest in the application of argumentation within the artificial agent societies. This paper extends the operational semantics to speech-act based communication messages received by an AgentSpeak(L) agent in order to enable argumentation in cognitive agents. The aim is to give semantics and implementation as logic-based plans for some key illocutionary forces, used for argumentation in the Belief-Desire-Intention (BDI) agent communication language 'AgentSpeak(L)'. The extension allows agents engaged in a dialogue to put forward their arguments, question beliefs of other agents more expressively. Therefore, using extended speech-act based communication; an agent can share its internal state with other agents and influence other agents' states. This work also provides a new dimension to argumentation based negotiation in BDI agents as this would enable the agents to negotiate using argumentation. Argumentation based negotiation can provide a powerful tool for the agents communicating to fix a deal using the electronic commerce services.

Keywords: Argumentation based negotiation, performatives, illocutionary forces, agents, AgentSpeak(L).

1 Introduction

The AgentSpeak(L) programming language was first introduced by Rao in 1996 [14]. It is based on logic programming and provides an elegant abstract framework for programming Belief-Desire-Intention (BDI) agents. AgentSpeak(L) is particularly interesting, in comparison to other agent-oriented languages, in that it retains the most important aspects of the BDI-based reactive planning systems on which it was based. AgentSpeak(L) has a working interpreter called Jason (introduced in 2005), and at the same time its formal semantics and relation to BDI logics are under constant research [4,5,6].

The previous work on giving operational semantics to AgentSpeak(L) [13, 15] account for the main illocutionary forces related to communicating AgentSpeak(L) agents. The interpreter described in that paper does not support argumentation expressively in communication as the kind of queries that the agent can put forward are limited. This is essential when it comes to engineering *multi-agent systems* for effective negotiation (bargain) and reasoning in electronic commerce services [1].

Negotiation using argumentation or argumentation based negotiation (ABN) allows agents to exchange additional information, or to “argue” about their beliefs and other mental attitudes during the negotiation process. In the context of negotiation, we view an *argument* as a piece of information that may allow an agent to: (a) *justify* its negotiation stance; or (b) *influence* another agent’s negotiation stance [10]. Therefore, using communication, an agent can share its internal state (beliefs, desires, intentions) with other agents and influence other agents’ states. Argumentation combined with BDI agents can provide a powerful tool, where agents are declarative, goal-based and capable of making plans. These agents can react to changes in environment and are able to reason about something that is uncertain or not acceptable. An environment where agents are not fully aware of their surroundings and possess uncertain and incomplete knowledge can benefit by communicating using argumentation.

This paper deals exactly with the speech-act based communication aspect of AgentSpeak(L), by extending its operational semantics to account for the main illocutionary forces related to the realization of argumentation in communicating AgentSpeak(L) agents. Our semantics tells exactly how to implement the processing of communication messages received by an AgentSpeak(L) agent (how its representation of Beliefs-Desires-Intentions is changed when a message is received). The concept of *plan* is used to simplify aspects of deliberation and knowing what course of action to take in order to achieve desired states of the world. Therefore, an AgentSpeak(L) agent (sender) *sends or utters* a speech-act based communication message whenever a special action for sending messages to other agents appears in the body of an intended plan that is being executed by the sender ‘s’. The important issue is then how to interpret a message that has been *received* by the receiver ‘r’. This is precisely the aspect of agent communication in argumentation that we consider in this paper.

In extending the operational semantics of AgentSpeak(L) to account for argumentation-based inter-agent communication, we also touch upon the issue of inclusion and interplay of various social factors like trust and social influence. To the best of our knowledge, our work is the first to give operational semantics incorporating the main illocutionary forces required for an argumentation based negotiation in a BDI programming language. We present a brief comparison of the proposed *extension in AgentSpeak(L)* to enable *argumentation* and henceforth ABN in the BDI agents with some of the existing work [16] focusing on argumentation based negotiation (ABN).

Comparing the proposed work with related work in ABN

Rahwan: He proposed interest based negotiation between BDI agents [17]. Our work is capable of handling arguments that reason about a decision as well as it has a provision by which trustworthy agents can suggest plan updates. Influential agents can demand penalty or send threats.

Ramchurn: They have worked on trust and persuasion in action-based agents only [18]. Our work will extend this further as we will be able to reason about the motives behind the agent’s goals and actions.

Jennings: Their work is also focused on action-based agents only. The latest contribution [11] deals with the issue of computational conflicts and argumentation in a social context for action-based agents only. Our work is capable of handling arguments that will take care of trust and power (influence) while agents communicate.

Parsons: Their work is based on logic and cooperative agents [16]. We are developing BDI logic-based agents who are cooperative but also self-interested to satisfy their own goals.

We are also in process of extending a BDI language so that, our proposals can be implemented and tested thoroughly.

The rest of the paper is organized as follows: followed by introduction, section 2 elaborates the various AgentSpeak(L) implemented KQML performatives or the illocutionary forces (ilfs) being used to enable argumentation amongst agents. It gives in detail the extended operational semantics for the various illocutionary forces (ilfs) and their implementation using Jason plans in brief. This section uses illustrative examples to show the arguments generated by each proposed ilfs for agent communication. Finally, section 3 demonstrates how the agents negotiate and argue using the proposed ilfs, to reach a desirable deal in electronic share trading. Section 4 presents conclusion and the future work.

2 Extending the Operational Semantics of AgentSpeak(L) for Speech-Act Based Communication between Agents

Presently, Jason makes available nine KQML – like performatives (communication messages send by the agents) [12]: tell, untell, achieve, unachieve, askone, askall, tellhow, untellhow and askhow for speech-act based communication between agents. These performatives are implemented as a set of AgentSpeak(L) plans in Jason. These plans have to be redefined in an agent source code (.asl) to include new features for agents to argue and negotiate in a multi agent society. Also new performatives are required to be implemented in the BDI language to enable argumentation.

The performatives existing for speech-act based communication needs to be overridden in agent's .asl program. Trust and influence are user defined SocAcc() and Power() methods of the Jason.as_semantics.agent class. All the performatives [7] are used with the internal action .send. The reason behind the particular manner of realization is that at the beginning of every reasoning cycle, agent checks for the messages they might have received from other agents. Any message received by them goes in a mailbox and is checked using the default checkmail method, which has the following structure:

<sender, illocutionary_force, content>,

where sender is the agent id of the message sender, *illocutionary_force* or *the ilf* is defined to be the intention of the sender that is represented by the used performative and content is the proposition send along with the ilf by the sender. The message send by the agent will take the following form:

.send(receiver_agent, ilf, propositional_content).

Therefore, a message is communicated using the .send internal action.

This work contributes towards the extension of existing speech-act based communication by defining new KQML performatives as illocutionary forces.

In the following sub-sections, we define and implement the following performatives:

advertise, unadvertise, ask-if, ask-if how, ask-all how, tell-all how, add-plan, remove-plan, threaten, penalize.

These are also theoretically grounded by giving operational semantics for them.

2.1 Defining Proposed Illocutionary Forces

Following are the ilfs that agents use to communicate while they are using argumentation for example, while negotiating to settle a deal in an electronic share trading scenario. The sender is represented by **s** and the recipient is **r**. KB is the agent's knowledge base and BB stands for its belief base.

advertise: **s** informs that it is particularly suited for processing a performative.

unadvertise: **s** informs that it is not suited for processing a particular performative, which it was capable to do earlier, so necessary changes is made in **r**'s KB.

ask-if: **s** wants to know if the sentence is in **r**'s KB.

ask-if how: **s** wants to know how that sentence exists in **r**'s BB, that is the reason and the source behind its existence.

ask-all how: **s** wants to know all the relevant plans of **r** behind all of **r**'s answers to a question. This also includes the sources behind those beliefs or answers.

tell-all how: **s** informs **r** about the particular plans in its own KB that lead to an event. If the source is trusted then **r** can add them in its own KB else it will ignore them. This performative is used to implement the ask-all how.

add-plan: **s** can make **r** include a plan in its library, if **s** is exercising strong social influence or power on **r**.

remove-plan: **s** can make **r** remove a plan in its library matching against a given literal, if **s** is has strong social influence on **r**.

threaten: **s** sends a declaration of an intention or determination to inflict punishment (warning) in retaliation for some action done on part of **r**.

penalize: **s** informs **r** that it is subjecting **r** to a penalty in terms of cost that it is supposed to pay.

Argumentation will allow these agents to reason behind each other's proposals. An agent can ask for reasons from another agent and know the motives behind a decision. Proposed ilfs allows agent to ask a question about other agent's belief (ask-if) ; to ask for an explanation using (ask-all how). Agents are able to present their arguments to another agent in the form of these ilfs while questioning or replying. This enables argumentation. The BDI agents will communicate and argue using them while negotiating in an electronic trade.

2.2 Informal Semantics of Receiving Messages

We here explain the illocutionary forces informally, considering an example of two agents, a buyer and a broker dealing in share trade. The broker informs the buyer

about the available shares for a particular company and price using `available_share` (Name, Price). We show how the two agents pass arguments using the `.send` action. The meaning of every argument is explained as well. Assume that the agent 's', the sender (in this case, broker agent) has executed the following formula that appeared in one of its intentions:

```
.send( r, advertise, available_share (Name, Price));
```

Agent 'r' (in this case, buyer agent) would then receive the given message in its mailbox,

```
<s, advertise, available_share (Name, Price)>
```

As soon as the message is taken from r's mailbox it is checked for trust and power if applicable on it. If the message is accepted by the agent then it is processed by agent 'r' as explained below. The list below gives various combinations of ilfs and the content types. Firstly, the internal action used by 's' to send the message is given and then the message is expressed in the form as it is received by the agent 'r'. We now consider each type of message (proposed in this work) that could be sent by 's' while arguing during negotiation. The examples are organized in the following categories: information exchange through advertisement, information seeking, know-how related communication, exercising influence and finally the warning and punishment messages.

- **Information exchange through advertisement**

.send(r, advertise, available_share(Name, Price)): This is the message sent by the sender 's' and the receiver 'r' receives following message in its mailbox `< s, advertise, available_share(Name, Price)>`. The belief `available_share(Name, Price)` [source(s)] will be added to r's belief base. As a result of receiving an advertisement the agent 'r' may also add a desire to send a tell message with content 'cfp' to the sender 's'. Content 'cfp' stands for call for proposal, which is sent by buyer to the broker so that broker can send its appropriate proposal accordingly.

.send(r, unadvertise, available_share(Name, Price)): The belief `available_share(Name, Price)` [source(s)] will be removed from r's belief base. Also the desire (if any) to send a 'cfp' will be dropped by 'r', if the cfp has not been send yet.

- **Information seeking**

.send(r, askif, share(Name, Price)): This is the argument sent by the sender and the receiver 'r' receives following message in its mailbox `< s, askif, share(Name, Price)>`. The sender 's' here, wants to know whether the content is true or not in the r's belief base. This argument triggers a test goal (+? content) in the mind of another agent. This can be useful whenever the buyer in share trading have to consult other trustworthy agents. As a result the receiver will reply back by sending a tell message with an answer true or false to the sender's query.

- **Know-how related communication**

.send(r, askallhow, "+! share(Name, Price); +! trust(AgentX, Task)"): This argument is sent by the agent 's' to 'r'. The receiver 'r' is supposed to answer back with the list of all relevant plans it currently has for the given particular triggering events in the argument. In this example, 's' wants 'r' to pass on his know-how reasons (plans) on how to achieve the given events in the string. The receiver will reply back using a `tellallhow` message that would be send to the sender.

.send(r, tellallhow, [“+! share(Name, Price): market(Name, Price) ← determine(Name, Price).”]; “+! trust(Broker, Supply): supply(AgentX, Price) ← determine(Broker, Trust).”]): This is sent by the sender to receiver in order to answer his argument about the plans that can be used to achieve a list of goals. Each member of the list sent in the message must be a string that can be parsed into a plan and all those plans are added to r’s plan library if the source is trusted.

- **Exercising influence**

.send(r, addplan, “@ determineprice +! share (Name, Price): market(Name, Price) ← determine(Name, Price).”): This argument allows an influential agent to make the receiver ‘r’ add a plan in its plan library.

.send(r, removeplan, “@ determineprice +! share (Name, Price): market(Name, Price) ← determine(Name, Price).”): This argument allows an agent to exercise its influence on ‘r’ by making it remove a plan. The string parsed in the message content can be a plan label or the plan itself sent by ‘s’.

- **Warning and Punishment messages**

.send(r, threat, [deceive (Price, Date), decrease_credit(AgentX)]): This is the argument sent by the sender and the receiver r receives following message in its mailbox < s, threat, [deceive(Price, Date), decrease_credit(AgentX)]>. The sender ‘s’ intend to threaten ‘r’, that if ‘r’ acts upon the goal of deceiving (on price and delivery date) then it will make ‘r’ infamous using decrease_credit and it will lose credit in the market. So, as a result if sender is influential enough then ‘r’ decides to drop the intention of deceiving ‘s’ and adds to its belief base that ‘s’ can decrease_credit(AgentX) [source(s)].

.send(r, penalty, [delayed (Offer), cost(Y)]): The sender asks the receiver to pay a penalty costing Y to him because of the reason specified as delayed (Offer). Therefore, then ‘r’ adds a goal of paying a cost.

2.3 Operational Semantics for Illocutionary Forces (KQML - Performatives)

This section presents the formal semantics of AgentSpeak(L), as originally given in [4]. The formal semantics is given in the style of Plotkin’s structural operational semantics, a widely used method for giving semantics to programming languages and studying their properties. Our present work is aligned with existing definitions for various ilfs, and we attempt to give a precise definition of the newly proposed speech–act based communication performatives required for argumentation. It defines and gives formal operational semantics of what it means for an AgentSpeak(L) agent to believe, desire, or intend a certain formula. Below, we give a brief description of an agent and its circumstance in a transition system [15].

An agent and its circumstance form a configuration of the transition system (an environment that keeps changing due to change in its circumstance) giving operational semantics to AgentSpeak(L). The transition relation:

$$\langle ag, C \rangle \rightarrow \langle ag', C' \rangle$$

is defined by the semantic rules given below.

An agent's circumstance C is a tuple $\langle I, E, M, A, R, Ap, i, \rho, \varepsilon \rangle$ where:

– I is a set of *intentions* $\{i, i', \dots\}$. Each intention i is a stack of partially instantiated plans.

E is a set of *events* $\{(te, i), (te', i'), \dots\}$. Each event is a pair (te, i) , where te is a triggering event and the plan on top of intention i is the one that generated te . The triggering event is a specific event for which a given plan is to be used from the plan library.

– M is a set which represents an agent's mail box. Messages are stored in a mail box and one of them is processed by the agent at the beginning of a reasoning cycle. The format of the messages stored in the mail box is $\langle Ilf, id, content \rangle$, where $Ilf \in \{advertise, unadvertise, ask-if, ask-if\ how, ask-all\ how, tell-all\ how, add-plan, remove-plan, threaten, penalize\}$ is the illocutionary force associated with the message, id identifies the agent that sent the message and $content$ is the message content, which can be either an atomic proposition (at) or a plan (p). The selection function is called S_M , and selects one particular message from M .

– Trust (id, at) is true if id identifies a trusted information source on the subject denoted by the atomic formula at . It is till now only used to decide whether received messages will be processed or not, without any reference to more complex notions of trust except some recent developments [9] to include cognitive trust [8] into the system. When the source is "trusted", the information source for a belief is acquired from communication. The agent simply discards messages arriving from untrusted agents. Information on beliefs of completely untrusted sources is not worth even keeping in the belief base.

– Power (id, at) is true if the agent has a subordination relation towards agent id in regards to the denotation of the atomic formula at .

– A is a set of *actions* to be performed in the environment.

– R is a set of *relevant plans* which have a matching triggering event to the plans in the plan library.

– Ap is a set of *applicable plans* which are relevant and the context part of these plans are consistent with the agent's belief base.

– Each circumstance C also has three components called ι , ε , and ρ . They keep record of a particular intention, event and applicable plan (respectively) being considered along the execution of an agent.

In order to keep the semantic rules neat, we adopt the following notations:

– If C is an AgentSpeak(L) agent circumstance, we write C_E to make reference to the component E of C . This holds true for all the other components of C .

– We write $C_i = \underline{\quad}$ (the underline symbol) to indicate that there is no intention being considered in the agent's execution. It is the same for C_ρ and C_ε .

– A metavariable is used, that is: *sources*, which is ranging over $\{percept, self, id\}$. These are the possible sources of a belief where *percept* means input is perceived from the environment, *self* represents the source is agent's own belief base and *id* represents the agent-id of the agent responsible for providing the information.

– bs denotes a set of beliefs (the agent's initial belief base)

– bs' denotes a set of beliefs after the received message is processed.

– ps denotes a set of plans (the agent’s plan library)

– ps' denotes a set of plans after the received message is processed.

All agents share the same ontological space and other terminologies as stated above.

2.3.1 Semantic Rules

We now extend the semantics to account for the processing of speech-act based messages received by an AgentSpeak(L) agent involved in communication with another agent. The proposed semantic rules are as follows.

Receiving an Advertise Message

$$\text{AdvertiseRec} \frac{S_M(C_M) = \langle \text{Advertise}, id, at \rangle}{\langle ag, C \rangle \rightarrow \langle ag', C' \rangle}$$

where: $C'_M = C_M - \{ \langle \text{Advertise}, id, at \rangle \}$
 $bs' \models \begin{cases} at[sources \cup \{id\}], & \text{if } bs \models at[sources] \\ at[id], & \text{otherwise} \end{cases}$
 $C'_I = C_I \cup \{ \langle + cfp[id], \top \rangle \}.$

The content of the message is added to the belief base in case it was not there previously. If the information is already there, the sender of the message is included in the set of sources giving accreditation to that belief. Simultaneously, an intention to send a call for proposal (cfp) to the sender of the advertisement is added in the receiver’s BB. Since the event $+cfp[id]$ is an external event (it is result of some action performed by another agent) therefore, it has no agent intention attached to it and an empty intention is represented using \top .

Receiving an Unadvertise Message

$$\text{UnadvertiseRec} \frac{S_M(C_M) = \langle \text{Unadvertise}, id, at \rangle}{\langle ag, C \rangle \rightarrow \langle ag', C' \rangle}$$

where: $C'_M = C_M - \{ \langle \text{Unadvertise}, id, at \rangle \}$
 $bs' \models at[sources - \{id\}], \text{ if } bs \models at[sources]$
 $bs' \not\models at[id], \text{ if } bs \models at[id],$
 $C'_I = C_I \cup \{ \langle - cfp[id], \top \rangle \}.$

Only the sender of the message is removed from the set of sources giving accreditation to the belief. In cases where the sender was the only source for that information, the belief is removed from the receiver’s belief base. The intention to send a cfp is also removed from BB.

Receiving an Askif Message

$$\text{AskifRec} \frac{S_M(C_M) = \langle \text{AskIf}, id, at \rangle \text{Trust}(id, at)}{\langle ag, C \rangle \rightarrow \langle ag', C' \rangle}$$

$$\text{where: } C'_M = C_M - \{ \langle \text{AskIf}, id, at \rangle \}$$

$$C'_M = \begin{cases} C'_M \cup \{ \langle \text{Tell}, id, at \rangle \}, & \text{if } +? \text{ at}[id] = \text{true} \\ C'_M \cup \{ \langle \text{Tell}, id, \sim at \rangle \}, & \text{otherwise} \end{cases}$$

$$C'_E = C_E \cup \{ \langle +? \text{ at}[id], \top \rangle \}.$$

AskIf results in generation of a test goal which is added to receiver agent's goals. If the answer is true then sender is informed that the belief *at* is true otherwise $\sim at$ holds true.

Receiving a TellAllHow Message

$$\text{TellAllHowRec} \frac{S_M(C_M) = \langle \text{TellAllHow}, id, \{p_{ij}\} \rangle \text{Trust}(id, at)}{\langle ag, C \rangle \rightarrow \langle ag', C' \rangle}$$

$$\text{where: } C'_M = C_M - \{ \langle \text{TellAllHow}, id, \{p_{ij}\} \rangle \}$$

$$ps' = ps \cup \{p_{ij}\}$$

$$C_i = _$$

AskAllHow results in the receiver sending a *TellAllHow* message to the sender to answer his query, where it informs the sender about all the relevant plans for *at*. This is used to inform receiver about multiple plans. Also no change in intention is being considered here.

Receiving an AskAllHow Message

$$\text{AskAllHowRec} \frac{S_M(C_M) = \langle \text{AskAllHow}, id, \{at_i\} \rangle \text{Trust}(id, at) \oplus \text{Power}(id, at)}{\langle ag, C \rangle \rightarrow \langle ag', C' \rangle}$$

$$\text{where: } C'_M = C_M - \{ \langle \text{AskAllHow}, id, \{at_i\} \rangle \}$$

$$C'_M = \begin{cases} C'_M \cup \{ \langle \text{TellAllHow}, id, \{p_{at}\} \rangle \}, \\ \text{and } p_{at} = \{ p_i \mid p_i \in \text{RelevantPlans}(at_i) \} \\ \text{no action} \text{ otherwise} \end{cases}$$

AskAllHow results in the receiver sending a *tellallhow* message to the sender to answer his query, where it informs the sender about all the relevant plans for at_i , that is $\{p_{at}\}$. If in case there are no plans then the receiver will not take any action. This performative can be useful in multi-issue negotiation where the agents have to reason about more than one related literals at_i during a particular time.

Receiving an AddPlan Message

$$\text{AddPlanRec} \frac{S_M(C_M) = \langle \text{AddPlan}, id, p \rangle \text{Trust}(id, at) \oplus \text{Power}(id, at)}{\langle ag, C \rangle \rightarrow \langle ag', C' \rangle}$$

where: $C'_M = C_M - \{ \langle \text{AddPlan}, id, p \rangle \}$
 $ps' = ps \cup \{p\}$
 $C_i = _$

Receiving a RemovePlan Message

$$\text{RemPlanRec} \frac{S_M(C_M) = \langle \text{RemovePlan}, id, p \rangle \text{Trust}(id, at) \oplus \text{Power}(id, at)}{\langle ag, C \rangle \rightarrow \langle ag', C' \rangle}$$

where: $C'_M = C_M - \{ \langle \text{RemovePlan}, id, p \rangle \}$
 $ps' = ps - \{p\}$.
 $C_i = _$

These ifls; addplan and removeplan can be used by agents to exercise their influence over other agents. These can be used by the sender to influence other agent's plans by sending them a particular plan and asking them to add or remove it. Also, there is no intention that is being considered in the agent's execution right now.

Receiving a Threat Message

$$\text{ThreatRec} \frac{S_M(C_M) = \langle \text{Threat}, id, at, at' \rangle \text{Power}(id, at)}{\langle ag, C \rangle \rightarrow \langle ag', C' \rangle}$$

where: $C'_M = C_M - \{ \langle \text{Threat}, id, at, at' \rangle \}$
 $C'_E = \begin{cases} C_E \cup \{ \langle \neg! at, \top \rangle \} & \text{if } \text{power}(id, at) > \delta \\ C_E & \text{otherwise} \end{cases}$
 $b'_s \models at' [id]$

The content at' of the message is added to the belief base as a threat from the source id . In case, if the influence of the sender agent is greater than a threshold value of δ , then the receiver will drop its goal of achieving at , that is delete the goal under threat. If the sender isn't influential enough it will record the threat but will not drop the desire of achieving the goal under threat.

Receiving a Penalty Message

$$\text{PenaltyRec} \frac{S_M(C_M) = \langle \text{Penalty}, id, at, ac \rangle}{\langle ag, C \rangle \rightarrow \langle ag', C' \rangle}$$

where: $C'_M = C_M - \{ \langle \text{Penalty}, id, at, ac \rangle \}$
 $C'_E = C_E \cup \{ \langle \neg! ac, \top \rangle \}$.

An agent can be penalized in person by another and made to pay a cost specified by the literal *ac*. As a result the receiver will pay the cost as it is now added as one of the goals (*ac*).

2.4 AgentSpeak(L) Plans for Illocutionary Forces

In the present section we give AgentSpeak(L) Jason [7] plans for some of the illocutionary forces (KQML performatives) for speech-act based communication between agents.

```
//Jason plans to handle the proposed KQML performatives
// Variables:
//   Sender:  the sender (an atom)
//   Content: content (typically a literal)
//   MsgId:   message id (an atom)
/* ---- advertise performatives ---- */
@kqmlReceivedAdvertise
+!kqml_received(Sender, advertise, Content, _)
    <- .add_desire(Content);
    .print("Intention added for cfp from", Sender).
/* ---- know-how performatives ---- */
@kqmlReceivedTellAllHow
+!kqml_received(Sender, tellallhow, Content, MsgId)
    <- !add_all_plans_received(Sender,Content).
@kqmlReceivedTellallList
+!add_all_plans_received(Sender, [H|T])
    <- .add_annot(H, source(Sender), CA);
    +CA;
    .add_plan(H, Sender)
    !add_all_plans_received(Sender,T).
@kqmlReceivedAskAllHow
+!kqml_received(Sender, askAllHow, Content, MsgId)
    <- .relevant_plans(Content, List As String);
    .send(Sender, tellallHow, List As String, MsgId).
/* ---- exercising influence performatives -- */
// instruct an agent to add any plan
```

```

@kqmlReceivedaddplan
+!kqml_received(Sender, addplan, Content, MsgId)
    <- .add_plan(Content, Sender).
/* ---warning and punishment performatives -- */
// receiving a threat
@kqmlReceivedThreat
+!kqml_received(Sender, threat, Content, MsgId)
    <- !process_threat_received(Sender,Content).
@kqml processthreatreceived
+! processthreatreceived (Sender, [H|T])
    <- .add_annot(H, source(Sender), CA);
    +CA;
    .drop_desire(T).

```

3 A Negotiation Dialogue Using Argumentation between Agents

Here, we demonstrate, how the agents can use the proposed illocutionary forces in the speech-act based communication during argumentation with help of an example mentioned in section 2.2. By extending the operational semantics of AgentSpeak(L) for speech-act based communication between agents and realizing the same in a Java-based interpreter Jason, we can enable the cognitive agents to reason and argue while they negotiate during decision making (selecting the best possible deal in an electronic trade environment where agents may not be having complete, updated or accurate information about the market). Through communication an agent can share its internal state with other agents. With communication using argumentation, an agent can also influence other agent's state. Eventually, this would affect the decision making process and agents will be able to take better, well informed decisions.

The dialogue demonstrated below is taken from an electronic share trading scenario where agents play three different roles. Agent 'A' is a broker, 'B' plays user or the buyer and 'C' is the consultant. C helps B in gathering initial information about the share market prices and the trust values of various brokers who may not be well known to the buyer B. This happens due to the incomplete and uncertain nature of knowledge in the real world. Thereafter, the agents A and B communicate to advertise the offer, accept or reject a proposed offer, and even to argue, reason, negotiate using the information seeking and know-how related communication arguments. Agents also try to exercise influence and give warnings or punishments to affect mental state of other agents. Eventually, the goal is to strike the most acceptable deal in the share trade using argumentation based negotiation.

The negotiation dialogue given below illustrates the outcome of a share deal between two agents, a buyer and a broker, as it would appear on the Jason console if the agents are executed as verbose with option 2. The societal factors are handled internally and few details have been curtailed to keep the AgentSpeak(L) code neat.

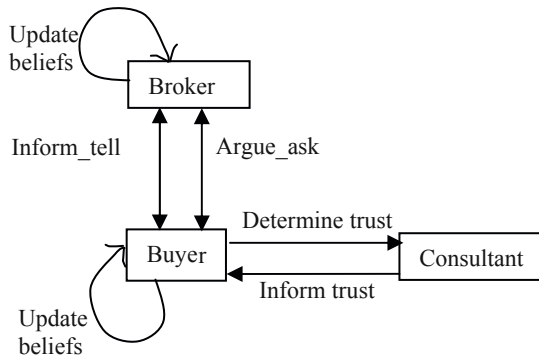


Fig. 1. A Negotiation Dialogue using Argumentation in Agents

This example illustrates argumentation based negotiation (ABN) when the buyer argues over the price of certain shares. The broker then gives its reasons for the price. Broker also asks buyer for the reasons it has for the quoted price that it is demanding. In the reasons generated by the buyer, the broker is able to find the buyer's other interests as well. Consequently, broker is able to present a new proposal to the buyer, which gets accepted by the buyer.

```

[broker] : adding belief plays(buyer, A) [source(percept)]
[broker] : added event +plays(buyer, A) [source(percept)]
[broker] : sending advertise share(X,Price) to buyer
[ buyer] : received message: <mid1,broker,advertise,
share(X, Price)>
[ buyer] : adding belief +share(X, Price) [source(broker)]
[ buyer] : adding belief +cfp(mid1,share(X, Reduceprice))
[source(percept)]
[ buyer] : added event +determinetrust(X, broker)
[ buyer] : sending askif trust(X, broker) to C
[ Consultant ]:added test goal +? trust(X, broker)
[Consultant]:sending tell trust(X,broker) to buyer
[ buyer] : adding belief + trust(X, broker)
[buyer] : sending cfp(mid2, share(X, Reduceprice))
[ buyer] : adding belief +share(X, Reduceprice)
// to remember my own proposal
[broker] : received message: <mid2, buyer, tell,
cfp (mid2, share(X, Reduceprice)>
  
```

```

[broker]:add belief +share(X,Reduceprice) [source(buyer)]
[broker]:sending askallhow +!share( , Reduceprice)
[buyer]:received message:<mid3, broker, askallhow,
+!share( _ , Reduceprice)>
[buyer]:send tellallhow["+!share(X, Reduceprice):
market (X, Reduceprice)←determine(X, Reduceprice)
source(C)]";"+! share( Y,Reduceprice)
← market(Y, Reduceprice)"]
[broker] : received message
[broker] : add plan "+! share( Y, Reduceprice)
← market(Y, Reduceprice)"]
[broker] : send tell propose(mid4,share(Y, Reduceprice))
[ buyer] : adding belief + share(Y, Reduceprice)
[ buyer] : sending askifhow share(Y, Reduceprice)
[broker] : sending tell source(share(Y, Reduceprice))
[ buyer] : sending threat [deceive(Price),
decrease_credit(broker)]
[broker] : adding belief + decrease_credit(self)
[source(buyer)]
[broker]: sending tell notdeceive(Price)
[ buyer] :sending tell
acceptproposal (mid4,share(Y,Reduceprice))
[ buyer]:adding belief + acceptproposal
( _ ,share(Y,Reduceprice))
[ buyer] : terminate session
[broker] : terminate session

```

As observed in the negotiation dialogue above, initially the buyer wanted to take the shares for 'X' at a reduced price. During argue and negotiation over the quoted price by the broker, the buyer receives another proposal of its interest from the broker to buy shares for 'Y' at a reduced price. Since the broker is trusted and buyer also makes sure that it will not be deceived, buyer agrees on the new proposal. Therefore, a satisfactory deal is fixed between the two agents at the end.

On the other hand, if argumentation was not used during negotiation by the agents then, it wasn't possible for the buyer to confirm the deal with the broker. There was also a possibility of failure of the deal due to lack of agreement between the two parties. Also, this argumentation resulted in an additional knowledge that the shares for 'Y' can also be purchased at a reduced price.

4 Conclusion and Future Work

ABN so far exists for action based agents only and some theoretical work has been done for the BDI agents as well. We propose to extend the concept of ABN to BDI (cognitive) agents by enabling argumentation in these agents. Implementing ABN for BDI agents would allow us to understand how the motives behind an agent's actions are modeled mentally and what effect argumentation can have on those motives to achieve certain actions. Argumentation can improve communication between the agents with increase in the number of interactions and knowledge transfer. This paper contributed towards the realization of argumentation as dialogue in a multi-agent negotiation, where agents are based on the BDI architecture. We achieved this by defining semantics and AgentSpeak(L) plans for the illocutionary forces that are required for communicating arguments between the agents. The semantics is given in the style of Plotkin's structural operational semantics, a widely used method for giving semantics to programming languages and studying their properties. The cognitive agents can now communicate to argue or even settle down for a deal amicably. A particular example from the share trading scenario is used to demonstrate the working of our proposed arguments in section 3.

Besides simple argumentation, it is also important to consider the computation of basic societal factors like trust and influence to negotiate efficiently in a trading scenario. Therefore, our future work will concentrate upon improving ABN in a multi-agent society by working on various argumentation strategies, conflict resolution and their interplay with the various societal factors like trust and social influence [3]. Work is in progress for developing trading and trust determination protocols [2] for the agents involved in argumentation. It is also possible to extend the communication between AgentSpeak(L) agents to a distributed SACI infrastructure. This will allow interfacing with agents created in other agent programming languages.

References

1. Bedi, P., Vashisth, P.: Negotiation using Argumentation for Location based E-Commerce in a Multi Agent Society. In: Proceedings of the International Conference on Artificial Intelligence (WORLDCOMP 2010), Las Vegas (2010a)
2. Bedi, P., Vashisth, P.: Designing Cognitive Agents for ABN in Electronic Trading. In: Proceedings of ICACCT 2010. IEEE, Haryana (2010b)
3. Bedi, P., Vashisth, P.: Social-Cognitive Trust Integrated in Agents for E-Commerce. In: Proceedings of IC4E 2011. IEEE, Mumbai (2011) (accepted for publication)
4. Bordini, R.H., Moreira, F.: Proving BDI properties of agent-oriented programming languages: the asymmetry thesis principles in AgentSpeak(L). *Annals of Mathematics and Artificial Intelligence*, Special Issue on Computational Logic in Multi-Agent Systems 42(1-3), 197–226 (2004)
5. Bordini, R.H., Braubach, L., Dastani, M., Seghrouchni, A.E.F., Gomez-Sanz, J.J., Leite, J., O'Hare, G., Pokahr, A., Ricci, A.: A survey of programming languages and platforms for multi-agent systems. *Informatica* 30(1), 33–44 (2006a)
6. Bordini, R.H., Hübner, J.F., Tralamazza, D.M.: Using Jason to implement a team of gold miners. In: Inoue, K., Satoh, K., Toni, F. (eds.) CLIMA 2006. LNCS (LNAI), vol. 4371, pp. 304–313. Springer, Heidelberg (2007)

7. Bordini, H., Hübner, J., Wooldridge, M.: Programing multi-agent systems in AgentSpeak using Jason. Wiley Series in Agent Technology (2007)
8. Castelfranchi, C., Falcone, R.: Social trust: A cognitive approach. In: Castelfranchi, C., Tan, Y.H. (eds.) *Trust and Deception in Virtual Societies*, pp. 55–90. Kluwer, Dordrecht (2001)
9. Hubner, J.F., Lorini, E., Vercouter, L., Herzig, A.: From cognitive trust theories to computational trust. In: *Proc. Of Workshop on Trust in Agent Societies (Trust@AAMAS 2009)*, pp. 55–67 (2009)
10. Jennings, N.R., Parsons, S., Noriega, P., Sierra, C.: On argumentation-based negotiation. In: *Proceedings of the International Workshop on Multi-Agent Systems*, Boston, pp. 1–7 (1998)
11. Karunatilake, N.C., Jennings, N.R., Rahwan, I., McBurney, P.: Dialogue Games that Agents Play within a Society. *Artificial Intelligence Journal* 173(9-10), 935–981 (2009)
12. Labrou, Y., Finin, T.: A semantics approach for KQML – a general purpose communication language for software agents. In: *Proceedings of the Third International Conference on Information and Knowledge Management (CIKM 1994)*. ACM Press, New York (1994)
13. Moreira, F., Vieira, R., Bordini, R.H.: Extending the operational semantics of a BDI agent-oriented programming language for introducing speech-act based communication. In: Leite, J., Zhang, S.-W., Sterling, L., Torroni, P. (eds.) *DALT 2003. LNCS (LNAI)*, vol. 2990, pp. 135–154. Springer, Heidelberg (2004)
14. Rao, A.S.: AgentSpeak(L): BDI agents speak out in a logical computable language. In: Perram, J., Van de Velde, W. (eds.) *MAAMAW 1996. LNCS*, vol. 1038, pp. 42–55. Springer, Heidelberg (1996)
15. Vieira, R., Moreira, A., Wooldridge, M., Bordini, R.H.: On the formal semantics of speech-act based communication in an agent-oriented programming language. *Journal of Artificial Intelligence Research* 29, 221–267 (2007)
16. Rahwan, I., Ramchurn, S.D., Jennings, N.R., McBurney, P., Parsons, S., Sonenberg, L.: Argumentation-based negotiation. *The Knowledge Engineering Review* 18(4), 343–375 (2003)
17. Rahwan, I.: Interest-based negotiation in multi-agent systems, PhD thesis, Dept. of Information Systems, University of Melbourne, Melbourne, Australia (2004)
18. Ramchurn, S.D.: Multi-agent negotiation using trust and persuasion, PhD thesis, School of Electronics and Computer Science, University of Southampton, U. K. (2004)

Medical Diagnosis Using Generic Intelligent Agents

Mukesh Kumar

University Institute of Engineering and Technology
Panjab University, Chandigarh
mukesh_rai9@yahoo.com

Abstract. Intelligent agents are a new paradigm for developing software applications. Intelligent Agent (IA) can be defined to be an autonomous, problem-solving computational entity capable of effective operation in dynamic and open environments. Intelligent Agents are often deployed in environments in which they can interact, can cooperate, with other agents, including both people and software. In this paper Intelligent Agent architecture is proposed. The proposed architecture is capable to perceive the dynamically changing environment. Based upon the perception, historical knowledge and internal design, the Intelligent Agent infers the action effecting the environment and updates the knowledgebase according to the action taken. Based upon the proposed Intelligent Agent architecture an Intelligent Medical Diagnosis System (IMDS) is given.

Keywords: Artificial Intelligence, Inference, Knowledgebase, Software Agent.

1 Introduction

Software agents have evolved from multi-agent systems (MAS), which in turn form one of three broad areas which fall under DAI, the other two being Distributed Problem Solving (DPS) and Parallel AI (PAI). Hence, as with multi-agent systems, they inherit many of DAI's motivations, goals and potential benefits. Software agents inherit DAI's potential benefits including modularity, speed (due to parallelism) and reliability (due to redundancy) [2]. It also inherits those due to AI such as operation at the knowledge level, easier maintenance, reusability and platform independence (Huhns & Singh, 1994). An agent can perceive its environment through sensors and act upon that environment through effectors [1]. As compared to a human agent which has sensors such as eyes, ears, hands and legs etc., a robotic agent substitutes cameras and infrared range finders for the sensors and various motors for the effectors. An agent's behavior depends only on its percept sequence to date.

An agent can be characterized by:

Autonomy: Software agents operate without the direct intervention of humans or other entities, and have some kind of control over their behaviors. They can execute automatically without any interaction with their environment. Autonomy is an essential attribute to distinguish software agents from general software procedures.

Proactivity: Traditional application procedures submissively accept users' commands to execute. Software agents can not only respond and react to their environment, but also they are able to take the initiative under certain situation.

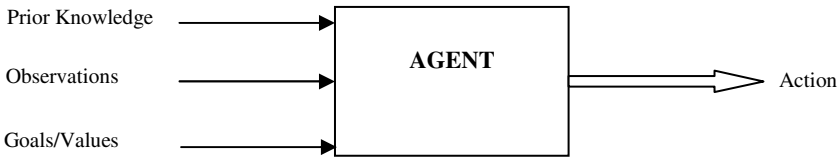


Fig. 1. General Agent

Reasoning: By learned knowledge and experience, software agents perform reasoning tasks in a rational way. Software agent intelligence comes from 3 major parts: the internal knowledge base, the self-adaptive ability and the knowledge-based reasoning ability.

Character: In the social activities, the factors that software agents need to consider are: security, risk and trust etc.

Communication/Cooperation/Coordination: These are the social attributes which reside in software agent groups.

Agent (see Fig. 1), can be described by making a table of the action it takes in response to each possible percept sequence. The agent perceives the environment using sensors. The agent stores everything it has perceived so far and a complete perceptual history of which makes up the percept sequence. The actions that the agent can perform depend upon the percept sequence. For the agent to respond efficiently and effectively, mapping functions are present. These mapping functions map the list of percept sequence to actions. A table of the action in response to each percept sequence is made.

(a)

<i>Percept Sequence</i>	<i>Action</i>
[A, Clean]	Right
[A, Dirty]	Suck
[B, Clean]	Left
[B, Dirty]	Suck
[A, Clean], [A, Clean]	Right
[A, Clean], [A, Dirty]	Suck

(b)

Fig. 2. (a) Vacuum Cleaner Agent Percepts: Location and Contents Action: Left, Right, Suck, NoOp, (b) Vacuum Cleaner Function

(For most agents, this would be a very long list—infinite, in fact, unless we place a bound on the length of percept sequences we want to consider an agent that operates on the basis of built-in assumptions will only operate successfully when those assumptions hold, and thus lack flexibility. So an Intelligent agent is required.

2 Related Work

The CogAff Architecture [3] of an intelligent agent consists of a central processing unit, perception and action. The central processing unit further consists of reflective

processes, deliberative reasoning reactive mechanisms. However, this architecture does not provide interagent communication that limits the use of software agents.

The RETSINA functional architecture [4] of an intelligent agent consists of four basic types of agents: interface agents, task agents, information agents and middle agents. This makes the overall structure of an intelligent agent complex, and the architecture is mainly used for Robotics software and not for simple application software.

BB1 [Hayes-Roth, 1985-present] [5] is an agent design for the *control problem*: which of its potential actions should a system perform next in the problem-solving process? In solving this problem, an agent system will decide what problems to solve, what knowledge to bring to bear, how to evaluate alternative solutions when problems are solved, and when to change its focus of attention. However BB1 does not consider adaptive ness an integral part of intelligent action (it does not purport to address intelligence, just the control problems), and it also makes an explicit distinction between domain and control activity.

In this paper a general architecture for intelligent agent is proposed. The proposed architecture is autonomous, proactive, collaborative, communicative, and adaptive to the environment.

3 The Proposed Intelligent Agent Architecture

The proposed intelligent Agent architecture is shown in Fig.3. In the present scenario: *An intelligent agent is an agent that percept its environment, normalizes perception, uses knowledgebase to infer the action, updates knowledgebase and affects the environment as according to the action.*

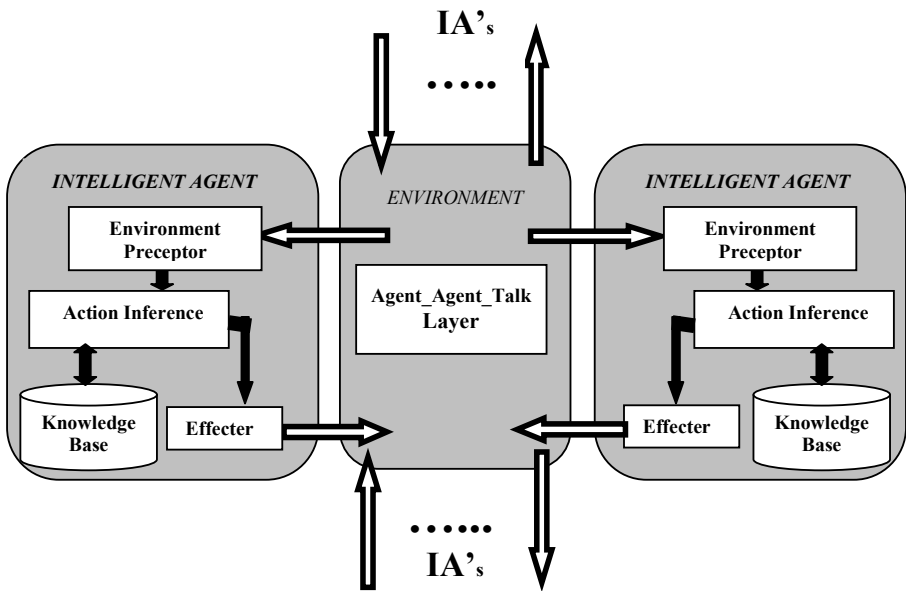


Fig. 3. Proposed Intelligent Agent Architecture

The following modules, from Fig.4 to Fig.8, explain the algorithmic behavior of the proposed Intelligent Agent:

INPUT for the IA: Environmental Perception.

OUTPUT from IA: Actions effecting the Current Environment

```

Intelligent_Agent ( Perception )
{
Normalized_Perception = Perception_Handler ( Perception ) ;

Action[ ] = Action_Inference(Normalized_Perception,
&Derived_Knowledge )

Update_KnowledgeBase( Derived_Knowledge ) ;

Environment_Effector (Action);
}

```

Fig. 4. Algorithm for Intelligent Agent

3.1 Inter Agent Communication (Agent_Agent_Talk Layer)

One of the most desirable features of an Agent is the inter-agent communication. The proposed architecture provides this feature with the help of Agent_Agent_Talk layer that exists in the current environment of the agent. Agent_Agent_Talk layer maintains a database containing the accessible agents, accessible rights of agent's data for each

```

Action[ ] Action_Inference
(Normalized_Perception,*Derived_Knowledge )
{
Action[1] = Inference derived using Normalized_Perception ,
Knowledge from

Knowledgebase and internal design of the Agent;

If (Need_To_Talk_To Other_Agent == 1)

Action[2] = ID_Of_Agent_To_Talk and Talk_Parameters;

Else

Action[2] = NULL;

Derived_Knowledge = Derive Knowledge constructs from the Action
taken;

Return (Action);
}

```

Fig. 5. Algorithm for Action Inference

```

Perception_Handler(Perception )
{
Normalized_Perception =Nomalize the Perception as required by
Action

Inference Mechanism;

Return(Normalized_Perception );
}

```

Fig. 6. Algorithm for Perception Handler

```

Environment_Effector (Action[ ])
{
Make the Environment to accept the changes as specified by the
Action[1];

If ( Action[2] != NULL )

    Invoke Agent_Agent_Talk layer with ID_Of_Agent_To_Talk,
and

    Talk_Parameters;
}

```

Fig. 7. Algorithm for Environment Effector

```

Update_KB( Derived_Knowledge )
{
Add Derived_Knowledge into Knowledgebase;
}

```

Fig. 8. Algorithm for Knowledgebase Updation

agent that exists in that environment. When *Action_Inference* decides that it should talk to some other agent, it tells the *Environment_Effector* to consult to the *Agent_Agent_Talk* layer and to communicate through some communication parameters passed by *Action_Inference*. It is the responsibility of the *Agent_Agent_Talk* layer to check for the Agent – Agent communication terms from the maintained database, and then accordingly to provide data access rights. The environment can be considered as the platform on which the agent is currently operating. The *Agent_Agent_Talk* layer is one of the threads of this platform.

4 Example: IMDS, Medical Diagnosis Using the Proposed Intelligent Agent Architecture

Intelligent agents can be used in medical diagnosis domain, if a patient is sitting away from the doctor then he will send a diagnosis request to Intelligent Medical Diagnosis

System (IMDS), IMDS will send an Intelligent Medical Diagnosis Agent (IMDA) to the patient's platform. IMDA will percept the symptoms from the patient's environment, normalizes the symptoms and send to the Action Inference. Action Inference with the help of normalized symptoms and the knowledgebase, which contains the previous knowledge in the form of facts and rules, infer the action to be taken and also according to the action also updates the knowledgebase. This action is communicated to the Environment Effector, who will tell the patient about the infected disease, precaution to be taken and various medicines for use. Also according to the symptoms and the action inferred IMDA will update the knowledgebase for further use. The complete medical diagnosis process using the proposed Intelligent Agent can be well understood with the help of following diagnosis instance:

Knowledgebase uses the rules and facts to represent the knowledge, for IMDS the rules are of the form:

- R1: If temp is High and Vomiting is High and Stomach_Pain = 1 and Headache is High Then Disease is TYPHOID.**
Action: Precautions + Medicines
- R2: If temp is High and Vomiting is High and Headache is High Then Disease is MALARIA.**
Action: Precautions + Medicines
- R3: If TDS = 4.85 and C-BLUE IS absent Then Disease is MELANOMA (Benign-nevus).**
Action: Precautions + Medicines
- R4: If TDS = 4.85 and C-BLUE IS present THEN the Disease is MELANOMA (Blue-nevus)**
Action: Precautions + Medicines
- R5: If TDS > 5.45 THEN the Disease is MELANOMA (Malignant)**
Action: Precautions + Medicines
- R6: If TDS > 4.85 and TDS < 5.45 THEN the Disease is MELANOMA (Suspicious)**
Action: Precautions + Medicines
- R7: If.....**
- R8: If..... and so on.**

Let the perceived symptoms from the patient are:

Body Temperature = 103⁰ F
Vomiting Level = .70
Stomach Pain = 1
Headache Level = .75

These symptoms are normalized by the Perception_Handler, means to say that Perception_Handler will normalize the symptoms to the data formats as required by the Action Inference mechanism. For example IMDS Perception_Handler will check whether 103⁰ F temperature lies in the High, Low, Very low or Very High temperature group, whether .75 , Headache Level corresponds to High, Very High or Low Headache and all these normalized results are passed to the Action Inference (Action_Inference). Action_Inference with the help of normalized perceptions and the

knowledgebase infers proper action to be taken. For this particular set of symptoms Rule 1 is fired and its action part is communicated to the Environment_Effector which will direct the patient for taking precautions and specified medicines.

Now let's consider some hypothetical symptoms and diseases for the Knowledgebase Updation (Adaptiveness).

Let two of the Knowledgebase rules are of the form:

Rule m: *If Symptom1 is A and Symptom2 is B and Symptom3 is C Then Disease is D1*
Action:[.....]

Rule n: *If Symptom1 is D1 and Symptom2 is E and Symptom3 is F Then Disease is D2*
Action: [.....]

And let the normalized perceived symptoms of a patient are A, B, C, E and F.

Now while inferring action, the Action_Inference will fire Rule m and diagnose the disease as D1 and after that it find that Rule n gets activated and diagnose the disease as D2 and communicate the Action part of rule n to the Environment_Effector. From the activation of Rule n the knowledgebase is updated by Update_Knowledge module which takes the derived knowledge that Symptoms A, B, C, E, F collectively can cause disease D2 and add it to the Knowledgebase.

Now let's consider the case of Agent-Agent communication. Let from the perceived symptoms IMDA infer that the patient should undergo an blood test and only after that it cal give diagnosis results, and for blood test there is an other agent IBTA (Intelligent Blood Test Agent). For this case IMDA will tell its Environment_Effector to consult the Agent_Agent_Talk layer with ID_IBTA, and Blood_Test_Parameters as parameters. Agent_Agent_Talk layer will determine from the maintained database whether IBTA could be invoked for that specification.

5 Conclusion

Intelligent Agent architecture is proposed. The proposed architecture is autonomous, proactive, collaborative, communicative, and adaptive to the environment. Inter-agent communication is provided by current environment layer. It avoids the redundant Agent Access information maintained by every Intelligent Agent because all this information is maintained centrally by the current environment layer. Based upon the proposed architecture an Intelligent Medical Diagnosis System (IMDS) is given. The functioning of IMDS as according to the proposed architecture is explained.

References

- [1] Artificial Intelligence: A Modern Approach by Stuart Russell and Peter Norvig. Prentice-Hall, Inc., Englewood Cliffs (1995)
- [2] Software Agents: An Overview Knowledge Engineering Review 11(3), 205–244 (October/November 1996)
- [3] How to think about architectures for human-like and other agents. Varieties of Evolvable Minds, Oxford (January 22, 2001)

- [4] The RETSINA Agent Architecture. The Software Agents Lab at Carnegie Mellon University's Robotics Institute, <http://www.cs.cmu.edu/>
- [5] <http://www.cs.dartmouth.edu/~brd/Teaching/AI/Lectures/Summaries/architectures>
- [6] Huang, J., Jennings, N.R., Fox, J.: An agent-based approach to health care management. *Int. Journal of Applied Artificial Intelligence* 9(4), 401–420 (1995)
- [7] A Concise Introduction to Multiagent Systems and Distributed AI Intelligent Autonomous Systems, Informatics Institute University of Amsterdam, Nikos Vlassis, University of Amsterdam (2003)
- [8] Noriega, P., Sierra, C. (eds.): AMET 1998 and AMEC 1998. LNCS (LNAI), vol. 1571. Springer, Heidelberg (1999)
- [9] Kim, D.S., Kim, C.S., Rim, K.W.: Modeling and Design of Intelligent Agent System. *International Journal of Control, Automation, and Systems* 1(2) (June 2003)
- [10] Cohen, P.R., Cheyer, A.J., Wang, M., Baeg, S.C.: An Open Agent Architecture. In: *Proceedings of AAAI Spring Symposium*, pp. 1–8 (1994)
- [11] Wooldridge, M., Jennings, N.: Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 115–152 (1995)
- [12] Genesereth, M., Ketchpel, S.: Software agents. *Communications of the ACM* 37, 48–53 (1994)
- [13] Finkelstein, A., Smolko, D.: Software Agent Architecture for Consistency Management in Distributed Documents. In: *Proceedings of Process support for Distributed Teambased Software Development Workshop, 4th World Multiconference SCI 2000/ISAS 2000, USA (2000)*

Computational Modeling and Dynamical Analysis of Genetic Networks with FRBPN- Algorithm

Raed I. Hamed^{1,2}

¹ Department of Computer Science, University of JMI, New Delhi-110025, India

² Department of Computer Science, Faculty of Computer, University of Anbar, Al-Anbar, Iraq
raed.inf@gmail.com

Abstract. Petri net is a well-established paradigm, where new algorithms with a sound biological understanding have been evolving. We have developed a new algorithm for modeling and analyzing gene expression levels. This algorithm uses fuzzy Petri net to transform Boolean network into qualitative descriptors that can be evaluated by using a set of fuzzy rules. By recognizing the fundamental links between Boolean network (two-valued) and fuzzy Petri net (multi-valued), effective structural fuzzy rules is achieved through the use of well-established methods of Petri net. For evaluation, the proposed technique has been tested using the nutritional stress response in *E.Coli* cells and the experimental results shows that the use of Fuzzy Reasoning Boolean Petri Nets (FRBPNs) based technique in gene expression data analysis can be quite effective and describe the dynamic behavior of genes.

Keywords: Gene expression levels, fuzzy logic, fuzzy Petri net, Boolean network.

1 Introduction

Computational models of genetic networks depend on the availability of data that reflect the state of the system. These data would ideally involve the expression rates of all genes plus the state of the types, concentrations and states of all proteins in the cell. Gene Regulatory networks represent genetic control mechanisms as directed graphs, in which genes are the nodes and the connecting edges signify regulatory interactions [1]. A review of modeling in genetic regulatory networks containing research on several techniques, such as differential equation [2, 3], fuzzy logic [4], Petri nets [5, 6], Boolean networks [7], Bayesian networks [8] and artificial neural networks [9]. The above-mentioned papers are dedicated to the applications of different methods to genetic networks and show that these methods are suitable to model special molecular biological systems.

The motivation for the development of the FPN model is to fuse the benefits of fuzzy logic (i.e. effectively manage uncertain or corrupted inputs, natural linguistic structure, etc.) with FPN techniques. The advantages of using FPNs in fuzzy rule-based reasoning systems include [11, 2]: (1) the graphical representation of FPNs model can help to visualize the inference states and modify fuzzy rule bases; (2) the analytic capability, which can express the dynamic behavior of fuzzy rule-based reasoning. Evaluation of markings is used to simulate the dynamic behavior of the system. The explanation of how to reach conclusions is expressed through the movements of tokens

in FPNs [11]. With this processing paradigm, FPN models are inherently able to represent and process uncertainty and imprecision information, things which are quite evident in all aspects of the real world problems. Logical and fuzzy Petri nets seem to be a good choice for knowledge representation and reasoning in GRNs, where the numerical values for the parameters characterizing the interactions and the concentrations are most often imprecise and fuzzy information.

In this paper, we modeled and analysed the genetic network by using fuzzy reasoning Boolean Petri nets and describe the dynamical behavior of gene. We illustrate our FRBPN approach by presenting a detailed case study in which the genetic regulatory network for the carbon starvation stress response in the bacterium *E. coli* [12] is modelled and analysed. Using the case study of data provided in [12] we define the Boolean behaviour of the key regulatory entities involved using truth tables. However, understanding the molecular basis of the transition between the exponential and stationary phase of *E. coli* cells has been the focus of extensive studies for decades [13].

The organization of this paper is as follows: In Section 2, the fundamental properties of aggregation operations and formal definition of fuzzy Petri nets are presented. In Section 3, we explain the details of the membership functions for active and inactive state with Boolean tables are investigated in this Paper. Section 4 describes the experimental results. Finally, we presented the conclusions of my model in Section 5.

2 Formal Definition of Fuzzy Petri Nets

The model was introduced by Looney [14] for the specification of rule-based reasoning using propositional logic. Places are interpreted as conditions having fuzzy truth values (tokens), while transitions represent the fuzzy decision values of rules. Reasoning in a FPN can be performed by iteratively maxing and mining transitions and fuzzy truth values of tokens. Formally, a fuzzy Petri net structure is defined as follows [10, 15]: The tuple $FPN = (P, T, D, I, O, F, \alpha, \beta)$ is called a fuzzy Petri net if:

1. $P = \{p_1, p_2, \dots, p_n\}$ is a finite set of places,
2. $T = \{t_1, t_2, \dots, t_n\}$ is a finite set of transitions, $P \cap T = \emptyset$,
3. $D = \{d_1, d_2, \dots, d_n\}$ is a finite set of propositions of FPRs. $P \cap T \cap D = \emptyset$, $|P| = |D|$, d_i ($i = 1, 2, \dots, n$) denotes the proposition that interprets fuzzy linguistic variables, such as: low, medium, high, as in my model;
4. $I: P \times T \rightarrow \{0, 1\}$ is an $n \times m$ input incidence matrix,
5. $O: P \times T \rightarrow \{0, 1\}$ is an $n \times m$ is an output incidence matrix,
6. $F = \{\mu_1, \mu_2, \dots, \mu_m\}$ where μ_i denotes the certainty factor (CF = μ_i) of R_i , which indicates the reliability of the rule R_i , and $\mu_i \in [0, 1]$,
7. $\alpha: P \rightarrow [0, 1]$ is the function which assigns a value between $[0, 1]$ to each place,
8. $\beta: P \rightarrow D$ is an association function, a bijective mapping.

Moreover, this model can be enhanced by including a function $Th: T \rightarrow [0, 1]$ which assigns a threshold value $Th(t_j) = \lambda_j \in [0, 1]$ to each transition t_j , where $j = 1, \dots, m$. Further more, a transition is enabled and can be fired in FPN models when values of tokens in all input places of the transition are greater than its threshold [2, 10].

In a fuzzy Petri net, different types of rules can be represented. **Type 1.** A simple fuzzy production rule. $I(t_j) = \{p_i\}$, $O(t_j) = \{p_i\}$, $f(t_j) = CF$, $\beta(p_i) = d_i$ and $\alpha(p_i) > 0$, where $1 \leq i, k \geq n$ and $1 \leq j \geq m$. It means that the degree of truth of the proposition $\beta(p_i) = d_i$ in this place p_i is equal $\alpha(p_i)$. **Type 2.** A composite conjunctive rule. $I(t_j) = P = \{p_1, p_2, \dots, p_l\}$, $O(t_j) = p_k$, $f(t_j) = CF$, $\beta(P) = [d_1, d_2, \dots, d_l]$ and $\alpha(P) = [\alpha(p_1), \alpha(p_2), \dots, \alpha(p_n)]$, where $1 \leq l, k \geq n$ and $1 \leq j \geq m$. **Type 3.** A composite disjunctive rule $I(t_j) = P = \{p_1, p_2, \dots, p_l\}$, $O(t_j) = p_k$, $f(t_j) = CF$, $\beta(P) = [d_1, d_2, \dots, d_l]$ and $\alpha(P) = [\alpha(p_1), \alpha(p_2), \dots, \alpha(p_n)]$, where $1 \leq l, k \geq n$ and $1 \leq j \geq m$.

3 Fuzzy Set for Active and Inactive State

In this section we consider the case of fuzzy set-based fuzzy model that is formed by using FRBPNs model. In this study, we carry out the modeling using characteristics of experimental data of a genetic regulatory network [12]. However, we present a method to illustrate how a FRBPNs model can be applied to get the optimal result. The method involves a genetic regulatory network with seven genes: Inputs Fis; CRP; Cya; Sig; GyrAB; TopA; SRNA, and outputs Fis>; CRP>; Cya>; GyrAB>; TopA>; SRNA>; Sig>. These set of genes are shown in Fig. 1, based on activation and inhibition relationships.

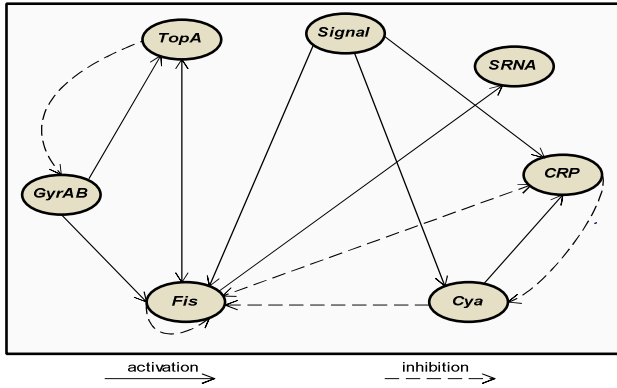


Fig. 1. Genetic network of *E. coli* entities includes both activation and inhibition relationships

The result of truth tables for each gene is shown by the following set of Boolean equations:

$$\begin{aligned}
 Cya &= \overline{Signal} + \overline{Cya} + \overline{CRP} & \overline{Cya} &= Signal & Cya &= \overline{CRP} \\
 CRP &= \overline{Fis} & \overline{CRP} &= Fis \\
 GyrAB &= (\overline{GyrAB} \ \overline{Fis}) + (\overline{TopA} \ \overline{Fis}) & \overline{GyrAB} &= (\overline{GyrAB} \ \overline{TopA}) + Fis \\
 TopA &= \overline{GyrAB} \ \overline{TopA} \ \overline{Fis} & \overline{TopA} &= \overline{GyrAB} + \overline{TopA} + \overline{Fis} \\
 Fis &= (\overline{Fis} \ \overline{Signal} \ \overline{GyrAB} \ \overline{TopA}) + (\overline{Fis} \ \overline{Cya} \ \overline{GyrAB} \ \overline{TopA}) + \\
 & \quad (\overline{Fis} \ \overline{CRP} \ \overline{GyrAB} \ \overline{TopA}) \\
 \overline{Fis} &= (\overline{CRP} \ \overline{Cya} \ \overline{Signal}) + Fis + \overline{GyrAB} + \overline{TopA} & \overline{SRNA} &= \overline{Fis}
 \end{aligned}$$

As an example, consider the truth table defining the behaviour of *TopA* shown in Table 1. We try to check the model is able to correctly switch between the exponential and stationary phases of growth control which is same as that in [12]. To make that the diagonal matrix B which can be got from the truth table for each gene in Table 1. Next the task is to construct the vector C_j . The entries of C_j correspond to change of probabilities of all the states when the matrix B is applied. Since we have the matrix B , the value of C_j can be constructed from B . The change of the probability of state j when the diagonal matrix for each gene is applied is the contribution of transitions from all the

Table 1. The truth table of each gene TopA

GyrAB	TopA	Fis	TopA
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	0

states. However, it is the sum of all the entries of the matrix B that denote transition to state j . To make the sum of all the entries in C_j , we need to use the following formula:

$$C(j) = \frac{1}{n} \sum_i [Dig(B)]_{i,j} \tag{1}$$

where n is the number of columns. As the vector $C(j)$ is determined, we need to calculate the finale value using equation 2. After the value C_j is determined, we need to use this value for each gene as input to FRBPNs model corresponding on the membership degree. As the value C_j is determined for each gene, these values help us to identify the procedure of fuzzy model by using it as initial parameters in our fuzzy model. In principle any function of the form $A : X \rightarrow [0, 1]$ describes a membership function associated with a fuzzy set A that depends not only on the concept to be represented, but also on the context in which it is used.

$$C_i = \sum_i^n C(j) \tag{2}$$

However, in our problem the membership functions with two linguistic value of active and inactive and with overlap 0.5 are plotted in Fig.2. However, these membership functions are used to make the decision on the gene behavior. With a numeric value g_i existing in a particular continuous variable's universe of discourse, we calculate a series of gene expressions to fuzzy sets existing in the frame of cognition of the variable. Conferring the character of Boolean Network, so the regulatory gene can be judged if it is active or inactive by function 3 as following:

$$G_i = \begin{cases} 0 \text{ (inactive)} & 0 \leq C_i < 0.35 \\ 1 \text{ (active)} & 0.35 \leq C_i \leq 1 \end{cases} \quad (3)$$

It is noted that the membership functions of Fig. 2 is either 1 or 0, singly depending on the truth of facts represented by C_i . For this task, membership functions Fig. 2 appear to be quite suitable, offering both flexibility and simplicity. The membership function is denoted as one-input–one-output so the rules will be as follows:

IF x is A_{j_1} THEN Z_{j_1} is inactive IF x is $A_{j_{i+1}}$ THEN $Z_{j_{i+1}}$ is active

The expression for the function defined by the fuzzy system when the input lies on the interval $x \in [a_{j_1}, a_{j_{i+1}}]$ will be:

$$f(x) = \frac{Z_{j_1} \mu_{j_1}(x) + Z_{j_{i+1}} \mu_{j_{i+1}}(x)}{\mu_{j_1}(x) + \mu_{j_{i+1}}(x)} \quad (4)$$

According to the system state represented by the place, the membership functions, designating the degree of truth, of the related fuzzy variables are determined.

4 Experimental Studies

Here we conduct a case study on the genetic regulatory network of the transition between the exponential and stationary phase of *E. coli* cells. It concerns a description of real estate in the *E. coli* which under normal environmental conditions, when nutrients are freely available, is able to grow rapidly entering an exponential phase of growth [13]. Our ultimate goal is to model the genetic regulatory network responsible for the carbon starvation nutritional stress response in *E. coli* cells based on the comprehensive data collated in [12]. The genes (*crp*, *cya*, *fis*, *gyrAB*, *topA*, and *rrn*) and their interactions believed to play a key role in the process, make up six modules, corresponding to the truth tables defining the Boolean behaviour of each regulatory entity in the nutritional stress response network for carbon starvation. Following the approach in [12], the level of *cAMP*, *CRP* and DNA supercoiling are not explicitly modelled as entities in my model. I initialise the model to a set of values representing the expression level of each gene. Starting from the initial conditions representing exponential growth the system is perturbed with an activate signal. The results expressed in table 2 show that the FRBPN approach provides an optimal completion (since results are the same as those observed by [12], [16]), while the model correctly switches from the exponential to the stationary phase of growth. To illustrate the analysis, we will focus on the fuzzy model at the Fis entity assuming CRP, Cya, Fis, GyrAB, TopA, and SRNA as inputs. The protein Fis is an important regulator of genes involved in the cellular growth and in addition, it controls the expression of CRP, as well as its own expression. The expression of the RNA gene is stimulated by the Fis protein. The level of RNAs is a reliable indicator of cellular growth.

Table 2. Set of quantified rules description of *E. Coli* growth

<i>If-condition</i>						<i>Then-condition</i>	<i>Confidence</i>
CRP	Cya	Fis	GyaAB	TopA	Sig	SRNA	Growth
High	High	Low					High 0.883
High	High						High 0.825
		Low	Medium	Low	High		Low 0.2
Medium	Medium	Low	Medium	Low	High		Low 0.120
		Low	Medium	Low	High		Low 0.101
	Low				High	Low	Low 0.01

5 Conclusion

In this paper we address the problem of combined fuzzy Petri net and Boolean networks to modeling and analyzing genetic regulatory network, from experimental data. I suggest using a fuzzy reasoning Boolean Petri nets (FRBPN) algorithm to identify the parameters of a system given by the observed data. Currently, only limited information on gene regulatory pathways is available in Systems Biology. The FRBPN algorithm apparently has higher computational complexity than general Boolean networks. The motivation for using fuzzy Petri nets models is the ability to translate Boolean data into linguistic constructs that can then be easily converted into testable hypotheses. It is also worth remarking that the quality values assigned by fuzzy Petri net to determine confidence values for cell growth in the *E.coli* are much more informative. We have shown here, that the FRBPN model is appropriate and can reach the same accuracy performance of available tool.

References

1. Weaver, D., Workman, C., Stormo, G.: Modeling regulatory networks with weight matrices. In: Pacific Symposium Biocomputing, vol. 99(4), pp. 112–123 (1999)
2. Hamed, R.I., Ahson, S.I.: Designing Genetic Regulatory Networks Using Fuzzy Petri Nets Approach. IJAC 7(3), 403–412 (2010)
3. Chen, T., He, H.L., Church, G.M.: Modeling gene expression with differential equations. In: Pacific Symposium on Biocomputing 1999, pp. 29–40 (1999)
4. Resson, H., Natarjan, P., Varghese, R.S., Musavi, M.T.: Applications of fuzzy logic in genomics. Journal of Fuzzy Sets and Systems 152, 125–138 (2005)
5. Matsuno, H., Doi, A., Nagasaki, M., Miyano, S.: Hybrid Petri net representation of gene regulatory network. In: Pacific Symposium on Biocomputing, vol. 5, pp. 338–349 (2000)
6. Matsuno, H., Fujita, S., Doi, A., Nagasaki, M., Miyano, S.: Towards Biopathway Modeling and Simulation. In: Proceedings of ICATPN, pp. 3–22 (2003)
7. Akutsu, T., Miyano, S., Kuhara, S.: Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In: Pacific Symposium on Biocomputing 1999, pp. 17–28 (1999)
8. Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian network. Bioinfo. 19, 2271–2282 (2003)

9. Vohradsky, J.: Neural networks model of gene expression. *The FASEB* 15, 846–854 (2002)
10. Hamed, R.I., Ahson, S.I.: A New Approach for Modeling Gene Regulatory Networks Using Fuzzy Petri Nets. *Journal of Integrative Bioinformatics* 7(113), 1–16 (2010)
11. Hamed, R.I., Ahson, S.I., Parveen, R.: Fuzzy Reasoning Boolean Petri Nets Based Method for Modeling and Analysing Genetic Regulatory Networks. In: Ranka, S., Banerjee, A., Biswas, K.K., Dua, S., Mishra, P., Moona, R., Poon, S.-H., Wang, C.-L. (eds.) *IC3 2010. CCIS*, vol. 94, pp. 530–546. Springer, Heidelberg (2010)
12. Ropers, D., de Jong, H., Page, M., Schneider, D., Geiselmann, J.: Qualitative Simulation of the Nutritional Stress Response in *E. coli*. *INRIA*, no. 5412 (2004)
13. Hengge-Aronis, R.: The general stress response in *Escherichia coli*. In: Storz, G., Hengge-Aronis, R. (eds.) *Bacterial Stress Responses*, pp. 161–178 (2000)
14. Looney, C.G.: Fuzzy petri nets for rule-based decision making. *IEEE Trans. Sys. Man and Cyb.* 18, 178–183 (1988)
15. Chen, S.M., Ke, J.S., Chang, J.F.: Knowledge Representation Using Fuzzy Petri Nets. *IEEE Transactions on Knowledge and Data Engineering* 2(3), 311–319 (1990)
16. Steggles, L.J., Banks, R., Wipat., A.: Modelling and Analysing Genetic Networks: From Boolean Networks to Petri Nets. In: Priami, C. (ed.) *CMSB 2006. LNCS (LNBI)*, vol. 4210, pp. 127–141. Springer, Heidelberg (2006)

A Petri Net-Fuzzy Predication Approach for Confidence Value of Called Genetic Bases

Raed I. Hamed^{1,2}

¹ Department of Computer Science, University of JMI, New Delhi-110025, India

² Department of Computer Science, Faculty of Computer, University of Anbar, Al-Anbar, Iraq
raed.inf@gmail.com

Abstract. We propose a novel solution for determining the confidence of a particular called genetic base within a sequence being called correctly. This has made measures of confidence of base call important and fuzzy methods have recently been used to approximate confidence by responding to data quality at the calling position. A fuzzy Petri net (FPN) approach to modeling fuzzy rule-based reasoning is proposed to determining confidence values for bases called in DNA sequencing. The proposed approach is to bring DNA bases-called within the framework of a powerful modeling tool FPN. The FPN components and functions are mapped from the different type of fuzzy operators of If-parts and Then-parts in fuzzy rules. The validation was achieved by comparing the results obtained with the FPN model and fuzzy logic using the MATLAB Toolbox; both methods have the same reasoning outcomes. Our experimental results suggest that the proposed models, can achieve the confidence values that matches, of available software.

Keywords: Confidence value, fuzzy logic, fuzzy Petri net algorithm.

1 Introduction

A major challenge of modeling biological systems is that conventional methods based on physical and chemical principles require data that is difficult to accurately and consistently obtain using either conventional biochemical or high throughput technologies, which typically yield noisy, semi-quantitative data (often in terms of a ratio rather than a physical quantity) [1]. Various kinds of models have been studied to express biological systems such as differential equations [2, 3], Boolean networks [4, 5], Petri Nets [6, 7, 8], Bayesian networks [9] and artificial neural networks [10]. The above-mentioned papers are dedicated to the applications of different methods to genetic networks and show that these methods are suitable to model special molecular biological systems.

Fuzzy Petri net (PN) is a successful tool for describing and studying information systems. Incorporating the fuzzy logic with Fuzzy Petri Nets has been widely used to deal with fuzzy knowledge representation and reasoning [12, 13, 14, 15]. It has also proved to be a powerful representation method for the reasoning of a rule-based system. Such an approach is appropriate for the case where a state of the modeled system corresponds to a marking of the associated FPN. The motivation for the development

of the FPN model is to fuse the benefits of fuzzy logic (i.e. effectively manage uncertain or corrupted inputs, natural linguistic structure, etc.) with FPN techniques.

The advantages of using FPNs in fuzzy rule-based reasoning systems include [14, 16]: (1) the graphical representation of FPNs model can help to visualize the inference states and modify fuzzy rule bases; (2) the analytic capability, which can express the dynamic behavior of fuzzy rule-based reasoning. Evaluation of markings is used to simulate the dynamic behavior of the system. The explanation of how to reach conclusions is expressed through the movements of tokens in FPNs [16].

The method presented in this paper develops a fuzzy Petri net model that can predict the confidence values for each base called in DNA sequencing. This approach here utilizes the information that is gathered at the base, for more information (see [17]). This includes information on the *height*, *peakness*, and *spacing* of the base under consideration and the next likely base. In order to validate my approach, we compare my method to the fuzzy logic toolbox of MATLAB. The comparison is made in terms of the confidence value measure of the bases called in DNA sequencing. The similarity that we have discovered is that they both have the same conclusions.

The organization of this paper is as follows: In Section 2, fuzzy Petri nets are described and the formulation of fuzzy sets and linguistic variables modeling are presented. In Section 3, we explain the details of case study and DNA bases called together with simulation method are investigated in this Paper. Finally, we presented the conclusions of my model in Section 4.

2 Fuzzy Petri Net Model

The Mamdani fuzzy inference system [19] was proposed as the first attempt to control a steam engine and boiler combination by a set of linguistic control rules obtained from experienced human operators. The output membership functions of Mamdani models are fuzzy sets, which can incorporate linguistic information into the model. The computational approach described in this paper is Mamdani fuzzy Petri net (MFPN) that is able to overcome the drawbacks specific to pure Petri nets.

Fuzzy models describing dynamic processes compute the states $x(t+1)$, at a time instant $t+1$, from the information of the inputs $x(t)$ and $u(t)$, at time instant t :

$$x(t+1) = f(x(t), u(t)) \quad (1)$$

where $f(\cdot)$ is a fuzzy model with the structure shown in Fig. 1. Input Layer (Layer 1), as shown in Eq. (2), no calculation is done in this layer. Each node, which corresponds to the inputs, $x(t)$ and $u(t)$, only transmits input value to the next layer directly. The certainty factor of the transitions in this layer is unity.

$$O_i^{(1)} = x(t), u(t) \quad (2)$$

where $x(t)$ and $u(t)$ are the expression value of the i^{th} gene at time instant t , and $O_i^{(1)}$ is the i^{th} output of layer 1. The values of the inputs, $x(t)$ and $u(t)$, and of the outputs, $x(t+1)$, can be assigned linguistic labels, e.g., 'low-expressed' (L), 'medium-expressed' (M), and 'high-expressed' (H). The output link of layer 2, represented as the membership value, specifies the degree to which the input value belongs to the

respective label. Linguistic rules can be formulated that connect the linguistic labels for $x(t)$ and $u(t)$ via an IF-part, called an antecedent of a rule and the THEN-part.

Any input value can be described through a combination of membership values in the linguistic fuzzy sets. In order to measure these input and output metadata universally, I normalize them into the same standard scale of [0, 1]. The values of linguistic variables are fuzzified to obtain the membership degree by membership function. For example, $\mu_{low_P_{called}}(0.35) = 0.5$, $\mu_{medium_P_{called}}(0.35) = 0.5$, means the value, 0.35 belongs to medium with confidence value (i.e. truth degree) of 50% while 50% belongs to low. That is, a 3-d membership vector for the fuzzy sets low, medium, and high corresponding to fuzzy peaknesses (P_{called}) is generated and is given by:

$$VP_{called} = [\mu_{low_P_{called}}, \mu_{medium_P_{called}}, \mu_{high_P_{called}}]^T, \quad (3)$$

Similarly, peaknesses (P_{2nd}), height (H_{called}), height (H_{2nd}), spacing (ΔS_{next}), and spacing ($\Delta S_{previous}$) are defined as:

$$\begin{aligned} VP_{2nd} &= [\mu_{low_P_{2nd}}, \mu_{medium_P_{2nd}}, \mu_{high_P_{2nd}}]^T, \\ VH_{called} &= [\mu_{vlow_H_{called}}, \mu_{low_H_{called}}, \mu_{medium_H_{called}}, \mu_{high_H_{called}}, \mu_{vhigh_H_{called}}]^T, \\ VH_{2nd} &= [\mu_{vlow_H_{2nd}}, \mu_{low_H_{2nd}}, \mu_{medium_H_{2nd}}, \mu_{high_H_{2nd}}, \mu_{vhigh_H_{2nd}}]^T, \\ V\Delta S_{next} &= [\mu_{low_Delta S_{next}}, \mu_{medium_Delta S_{next}}, \mu_{high_Delta S_{next}}]^T, \\ V\Delta S_{previous} &= [\mu_{low_Delta S_{previous}}, \mu_{medium_Delta S_{previous}}, \mu_{high_Delta S_{previous}}]^T, \end{aligned}$$

The output of each node represents the firing strength of the corresponding fuzzy rule. The output of the node in layer 3 carry out min-processing (i.e., computing minimum values) in fuzzy inference, and output of the node in layer 4 carry out max processing (i.e., computing maximum values) in fuzzy inference. Lastly we need to find the crisp output of layer 5 by finding the *center of gravity* method as the defuzzification using *center of gravity* of the output of the node in layer 4.

3 Experimental and Simulation Results

As shown in the fuzzy rule base of the reasoning process as a part in the main system to determine the confidence value is constructed of three models. Here we describe existing variables (peakness, height, and spacing) following the Method [17] which has been used for comparative analysis. The fuzzy membership functions of these input variables are described. We input a crisp data (i.e. P_{called} , P_{2nd} , H_{called} , H_{2nd} , ΔS_{next} and $\Delta S_{previous}$) into those corresponding membership functions, and get the membership degree for all variables as listed in the fourth column of Table 1. To explain our method a part of a DNA sequence that involves six bases (ATCTCG) is presented as listed in the third column of Table 1. Table 1 shows the P_{called} , P_{2nd} , H_{called} , H_{2nd} , ΔS_{next} and $\Delta S_{previous}$ for the six bases. For example, for the base G the normalized value for each input data as follows: $P_{called} = 1$, $P_{2nd} = 0.62$; $H_{called} = 0.98$, $H_{2nd} = 0.49$; $\Delta S_{next} = 0.28$, and $\Delta S_{previous} = 0.3$. The membership degrees of these input data are calculated by Trapezoidal membership functions. These membership function value can be used as the truth degree of each antecedent proposition in our FPN models. For example, with base G the truth degree of the proposition listed as: For each input data the firing strength of each activated rule is calculated by the *MIN* and *MAX* composition operator, respectively. It yields Low: $MAX(FR_1, FR_2, FR_3, FR_6) =$

<i>Peakness base G</i>	<i>Height base G</i>	<i>Spacing base G</i>	<i>PEAKNESS BASE G</i>
$\mu_{\text{Flat}}(P_{\text{called}}) = 0$	$\mu_{\text{VLow}}(H_{\text{called}}) = 0$	$\mu_{\text{Small}}(\Delta S_{\text{next}}) = 1$	$\text{FR}_1 : \text{MIN}(0, 0) = 0,$
$\mu_{\text{Medium}}(P_{\text{called}}) = 0$	$\mu_{\text{Low}}(H_{\text{called}}) = 0$	$\mu_{\text{Medium}}(\Delta S_{\text{next}}) = 0$	$\text{FR}_2 : \text{MIN}(0, 0.8) = 0,$
$\mu_{\text{Sharp}}(P_{\text{called}}) = 1$	$\mu_{\text{Medium}}(H_{\text{called}}) = 0$	$\mu_{\text{Large}}(\Delta S_{\text{next}}) = 0$	$\text{FR}_3 : \text{MIN}(0, 0.2) = 0,$
$\mu_{\text{Flat}}(P_{2\text{nd}}) = 0$	$\mu_{\text{High}}(H_{\text{called}}) = 0$	$\mu_{\text{Small}}(\Delta S_{\text{previous}}) = 1$	$\text{FR}_4 : \text{MIN}(0, 0) = 0,$
$\mu_{\text{Medium}}(P_{2\text{nd}}) = 0.8$	$\mu_{\text{VHigh}}(H_{\text{called}}) = 1$	$\mu_{\text{Medium}}(\Delta S_{\text{previous}}) = 0$	$\text{FR}_5 : \text{MIN}(0, 0.8) = 0,$
$\mu_{\text{Sharp}}(P_{2\text{nd}}) = 0.2$	$\mu_{\text{VLow}}(H_{2\text{nd}}) = 0$	$\mu_{\text{Large}}(\Delta S_{\text{previous}}) = 0$	$\text{FR}_6 : \text{MIN}(0, 0.2) = 0,$
	$\mu_{\text{Low}}(H_{2\text{nd}}) = 0.1$		$\text{FR}_7 : \text{MIN}(1, 0) = 0,$
	$\mu_{\text{Medium}}(H_{2\text{nd}}) = 0.9$		$\text{FR}_8 : \text{MIN}(1, 0.8) = 0.8,$
	$\mu_{\text{High}}(H_{2\text{nd}}) = 0$		$\text{FR}_9 : \text{MIN}(1, 0.2) = 0.2,$
	$\mu_{\text{VHigh}}(H_{2\text{nd}}) = 1$		

$\text{MAX}(0, 0, 0, 0) = 0$, Medium: $\text{MAX}(\text{FR}_5, \text{FR}_9) = \text{MAX}(0, 0.2) = 0.2$, High: $\text{MAX}(\text{FR}_4, \text{FR}_7, \text{FR}_8) = \text{MAX}(0, 0, 0.8) = 0.8$, *Height base G*

$\text{FR}_1 : \text{MIN}(0, 0) = 0,$	$\text{FR}_{14} : \text{MIN}(0, 0) = 0,$	<i>Spacing base G</i>
$\text{FR}_2 : \text{MIN}(0, 0.1) = 0,$	$\text{FR}_{15} : \text{MIN}(0, 0) = 0,$	$\text{FR}_1 : \text{MIN}(1, 1) = 1,$
$\text{FR}_3 : \text{MIN}(0, 0.9) = 0,$	$\text{FR}_{16} : \text{MIN}(0, 0) = 0,$	$\text{FR}_2 : \text{MIN}(1, 0) = 0,$
$\text{FR}_4 : \text{MIN}(0, 0) = 0,$	$\text{FR}_{17} : \text{MIN}(0, 0.1) = 0,$	$\text{FR}_3 : \text{MIN}(1, 0) = 0,$
$\text{FR}_5 : \text{MIN}(0, 0) = 0,$	$\text{FR}_{18} : \text{MIN}(0, 0.9) = 0,$	$\text{FR}_4 : \text{MIN}(0, 1) = 0,$
$\text{FR}_6 : \text{MIN}(0, 0) = 0,$	$\text{FR}_{19} : \text{MIN}(0, 0) = 0,$	$\text{FR}_5 : \text{MIN}(0, 0) = 0,$
$\text{FR}_7 : \text{MIN}(0, 0.1) = 0,$	$\text{FR}_{20} : \text{MIN}(0, 0) = 0,$	$\text{FR}_6 : \text{MIN}(0, 0) = 0,$
$\text{FR}_8 : \text{MIN}(0, 0.9) = 0,$	$\text{FR}_{21} : \text{MIN}(1, 0) = 0,$	$\text{FR}_7 : \text{MIN}(0, 1) = 0,$
$\text{FR}_9 : \text{MIN}(0, 0) = 0,$	$\text{FR}_{22} : \text{MIN}(1, 0.1) = 0.1,$	$\text{FR}_8 : \text{MIN}(0, 0) = 0,$
$\text{FR}_{10} : \text{MIN}(0, 0) = 0,$	$\text{FR}_{23} : \text{MIN}(1, 0.9) = 0.9,$	$\text{FR}_9 : \text{MIN}(0, 0) = 0,$
$\text{FR}_{11} : \text{MIN}(0, 0) = 0,$	$\text{FR}_{24} : \text{MIN}(1, 0) = 0,$	
$\text{FR}_{12} : \text{MIN}(0, 0.1) = 0,$	$\text{FR}_{25} : \text{MIN}(1, 0) = 0,$	
$\text{FR}_{13} : \text{MIN}(0, 0.9) = 0,$		

$\text{VLow} : \text{MAX}(\text{FR}_2, \text{FR}_3, \text{FR}_4, \text{FR}_5, \text{FR}_7, \text{FR}_8, \text{FR}_9, \text{FR}_{10}, \text{FR}_{13}, \text{FR}_{14}, \text{FR}_{15}, \text{FR}_{19}, \text{FR}_{20}, \text{FR}_{25}) = \text{MAX}(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) = 0$, Low: $\text{MAX}(\text{FR}_1, \text{FR}_6, \text{FR}_{12}, \text{FR}_{18}, \text{FR}_{24}) = \text{MAX}(0, 0, 0, 0, 0) = 0$, Medium: $\text{MAX}(\text{FR}_{11}, \text{FR}_{17}) = \text{MAX}(0, 0) = 0$, High: $\text{MAX}(\text{FR}_{16}, \text{FR}_{23}) = \text{MAX}(0, 0.9) = 0.9$ $\text{VHigh} : \text{MAX}(\text{FR}_{21}, \text{FR}_{22}) = \text{MAX}(0, 0.1) = 0.1$ Low: $\text{MAX}(\text{FR}_6, \text{FR}_8, \text{FR}_9) = \text{MAX}(0, 0, 0) = 0$, Medium: $\text{MAX}(\text{FR}_3, \text{FR}_5, \text{FR}_7) = \text{MAX}(0, 0, 0) = 0$, High: $\text{MAX}(\text{FR}_1, \text{FR}_2, \text{FR}_4) = \text{MAX}(1, 0, 0) = 1$,

According to the result of *max* composition operation the defuzzification of output is used to make a final decision. We adopt the “center of gravity” method in [25] to solve this problem. Then, the defuzzification of peakness, height, and spacing is calculated as Peakness = 0.76, Height = 0.76, and Spacing = 0.82 by the centroid of the aggregate output membership function in the each FPNs model. Following the steps of the reasoning process, the final winning rule in Peakness FPN model is FR_8 (*IF P_{called} is Sharp and $P_{2\text{nd}}$ is Medium THEN the Peakness is High*), which indicates that the “Peakness is High”, in the Height FPN Model the final winning rule is FR_{23} (*IF H_{called} is VeryHigh and $H_{2\text{nd}}$ is Medium THEN the Height is High*), which indicates the “Height is High”, and in the Spacing FPN Model the final winning rule is FR_1 (*IF ΔS_{next} is Small and $\Delta S_{\text{previous}}$ is Small THEN the Spacing is High*), which indicates the “Spacing is High”. These peakness, height, and spacing values are then imported to

the antecedent propositions in the main system model to determine the confidence value. The fuzzy rules of main system are aggregated and defuzzied to have a crisp value of confidence value = 0.75. By calculating the centroid, which indicates the rule FR_{42} (*IF Peakness is High and Height is High and Spacing is High THEN the Confidence value is High*) is the winner.

Table 1. Confidence values for the bases called

	A	T	C	T	C	G
Peakness	0.819	0.826	0.5	0.5	0.569	0.76
Height	0.75	0.925	0.192	0.75	0.35	0.761
Spacing	0.826	0.826	0.826	0.826	0.826	0.826
Confidence	0.75	0.925	0.124	0.75	0.35	0.75

4 Conclusion

DNA sequence basecalling is at the heart of modern genomics, which is already contributing to healthcare innovation. In this paper, introduce a FPN model for fuzzy rule based reasoning. The fuzzy set theory and the fuzzy production rule method are used to establish the fuzzy rules for the confidence value prediction of the bases called in DNA sequencing. This includes the transformation of fuzzy rules into FPN, together with their reasoning. The motivation for using fuzzy Petri nets models is the ability to translate numeric data into linguistic constructs that can then be easily converted into testable hypotheses. It is also worth remarking that the quality values assigned by fuzzy Petri net to determine confidence values for bases called in DNA sequencing are much more informative. We have shown here, that the FPN model is appropriate and can reach the same accuracy performance of available software. The validation was achieved by comparing the results obtained with the FPN model and fuzzy logic using the MATLAB Toolbox; both methods have the same reasoning outcomes. It verifies that the confidence value of the bases called can be successfully reasoned by the proposed FPN model.

References

1. Fitch, J., Sokhansanj, B.: Genomic engineering moving beyond DNA sequence to function. *Proc. IEEE* 88, 1949–1971 (1971)
2. Novak, B., Csikasz-Nagy, A., Gyorffy, B., Chen, K., Tyson, J.: Mathematical model of the fission yeast cell cycle with checkpoint controls at the G1/S, G2/M and metaphase/anaphase transitions. *Biophysical Chemistry* 72, 185–200 (1998)
3. Chen, T., He, H.L., Church, G.M.: Modeling gene expression with differential equations. In: *Pacific Symposium on Biocomputing 1999*, pp. 29–40 (1999)
4. Liang, S., Fuhrman, S., Somogyi, R.: REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In: *Pacific Symposium on Biocomputing*, vol. 3, pp. 18–29 (1998)

5. Akutsu, T., Miyano, S., Kuhara, S.: Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In: Pacific Symposium on Biocomputing 1999, pp. 17–28 (1999)
6. Matsuno, H., Doi, A., Nagasaki, M., Miyano, S.: Hybrid Petri net representation of gene regulatory network. In: Pacific Symposium on Biocomputing, vol. 5, pp. 338–349 (1999)
7. Matsuno, H., Fujita, S., Doi, A., Nagasaki, M., Miyano, S.: Towards Biopathway Modeling and Simulation. In: Proceedings of ICATPN, pp. 3–22 (2003)
8. Fujita, S., Matsui, M., Matsuno, H., Miyano, S.: Modeling and simulation of fission yeast cell cycle on hybrid functional Petri net. IEICE Transactions on Fundamentals of Electronics, CCS E87-A(11), 2919–2928 (2003)
9. Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian network. *Bioinform.* 19, 2271–2282 (2003)
10. Vohradsky, J.: Neural networks model of gene expression. *The FASEB Journal* 15, 846–854 (2002)
11. Goss, P.J.E., Peccoud, J.: Analysis of the stabilizing effect of Rom on the genetic network controlling ColE1 plasmid replication. In: Pacific Sym. on Bioc., pp. 65–76 (1999)
12. Hamed, R.I., Ahson, S.I.: A New Approach for Modeling Gene Regulatory Networks Using Fuzzy Petri Nets. *Journal of Integrative Bioinformatics* 7(1), 1–16 (2010)
13. Hamed, R.I., Ahson, S.I.: Designing Genetic Regulatory Networks Using Fuzzy Petri Nets Approach. *IJAC* 7(3), 403–412 (2010)
14. Chen, S.M., Ke, J.S., Chang, J.F.: Knowledge Representation Using Fuzzy Petri Nets. *IEEE Transactions on Knowledge and Data Engineering* 2(3), 311–319 (1990)
15. Hamed, R.I., Ahson, S.I.: Fuzzy Reasoning Boolean Petri Nets Based Method for Modeling and Analysing Genetic Regulatory Networks. In: Ranka, S., Banerjee, A., Biswas, K.K., Dua, S., Mishra, P., Moona, R., Poon, S.-H., Wang, C.-L. (eds.) IC3 2010. CCIS, vol. 94, pp. 530–546. Springer, Heidelberg (2010)
16. Hamed, R.I., Ahson, S.I., Parveen, R.: From Fuzzy Logic Theory to Fuzzy Petri Nets Predicting Changes in Gene Expression Level. In: International Conference on Methods and Models in Computer Science, December 14–15, pp. 139–145 (2009)
17. Resson, H., Natarjan, P., Varghese, R.S., Musavi, M.T.: Applications of fuzzy logic in genomics. *Journal of Fuzzy Sets and Systems* 152, 125–138 (2005)
18. Qu, W., Shirai, K.: Belief learning in certainty factor model and its application to text categorization. In: Proceedings of the 2003 Joint Conference of the Fourth Inter. Con. on Infor., Comm. and Signal Processing, vol. 12, pp. 1192–1196 (2003)
19. Mamdani, E.H., Assilian, S.: An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies* 7(1), 1–13 (1975)
20. Zadeh, L.A.: The concept of linguistic variable and its applications to approximate reasoning-II. *Inform. Sci.* 8, 301–357 (1975)
21. Zadeh, L.A.: Precisiated natural language – toward a radical enlargement of the role of natural languages in information processing, decision and control. In: Proceedings of the Ninth International Conference on Neural Information Processing, vol. 1, pp. 1–3 (2002)
22. Berno, A.: A graph theoretic approach to the analysis of DNA sequencing data. *Genome Res.* 6(2), 80–91 (1996)
23. Human Genome Project Information, <http://www.ornl.gov/hgmis/>
24. Ewing, B., Hillier, L., Wendl, M., Green, P.: Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8, 175–185 (1998)
25. Negnevitsky, M.: *Artificial Intelligent—A Guide to Intelligent Systems*. Addison-Wesley, New York (2002)

Textural Feature Based Image Classification Using Artificial Neural Network

Salavi Rashmi and Sohani Mandar

Department of Computer Engg.,
Vidyalankar Institute of Technology, Wadala,
Mumbai - India
rashmisalvi@gmail.com,
mgsohani@rediffmail.com

Abstract. Image classification plays an important part in the fields of Remote sensing, Image analysis and Pattern recognition. Digital image classification is the process of sorting all the pixels in an image into a finite number of individual classes. The conventional statistical approaches for land cover classification use only the gray values. However, they lead to misclassification due to strictly convex boundaries. Textural features can be included for better classification but are inconvenient for conventional methods. Artificial neural networks can handle non-convex decisions. The uses of textural features help to resolve misclassification. This paper describes the design and development of a hierarchical network by incorporating textural features. The effect of inclusion of textural features on classification is also studied.

1 Introduction

A huge amount of digital images are added to the databases. It is inefficient to retrieve an image from a massive collection of image database. So it is necessary to make the database led to the proliferation of emerging storage. Image classification plays an important role in the field of pattern recognition. Instead of searching an image in huge databases, images are classified in classes and retrieved from that particular class.

The conventional statistical approach for image classification uses only gray values. However it leads to misclassification due to strictly convex boundary. Texture feature based image classification gives better classification. Artificial Neural network can handle non-convex decisions.

In texture classification the goal is to assign an unknown sample image to one of a set of known texture classes. Textural features can be either, scalar numbers, discrete histograms or empirical distributions. They characterize the textural properties of the images, such as spatial structure, contrast, roughness, orientation, etc and have some correlation with the desired output. There are fourteen textural features. The design considers four features namely angular second moment (ASM), contrast, correlation, variance. However, it can be extended by inclusion of all features.

ANNs can provide suitable solutions for problems, which are generally characterized by non-linearity's, high dimensionality noisy, complex, imprecise, imperfect or error prone sensor data, and lack of a clearly stated mathematical solution or algorithm.

A key benefit of neural networks is that a model of the system or subject can be built just from the data. Supervised learning is a process of training a neural network with examples of the task to learn, ie, learning with a teacher. Unsupervised learning is a process when the network is able to discover statistical regularities in its input space and automatically develops different modes of behavior to represent different classes of inputs.

The project depicts with huge amount of image database like medical images, satellite images, etc. and training the network to prepare training data. Using training data, Neural Network classifies the images into different classes based on their textural features extracted from co-occurrence matrix.

Image classification plays an important role in the field of pattern recognition. Instead of searching an image in huge databases, images are classified in classes and retrieved from that particular class. This work can also be used in remote sensing, Detective agencies.

2 Methodology

Image classification based on textural feature using Artificial Neural Network requires two steps:

2.1 Textural Feature Extraction Using Gray Level Co-occurrence Matrix

The GLCM, is a matrix of frequencies at which two pixels, separated by a certain vector occur in the image. The distribution in the matrix will depend on the angular and distance relationship between pixels.

GLCM features are extracted for ' $n \times n$ ' primitive template matrix in the directions 0° , 45° , 90° , and 135° , and then averaging is done to make them direction invariant.

2.2 Training Neural Network for Classification

This work uses three layered Neural Network that is Input, hidden, and output Layers. The features extracted, are fed as a input to the input layer of Neural Network and train the data. The data is propagated to the hidden layer and then to the output layer.

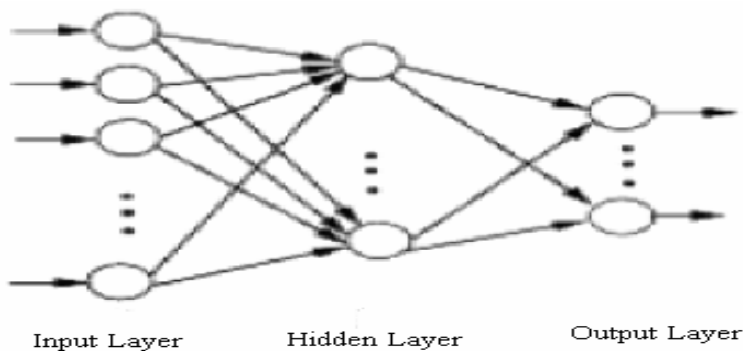


Fig. 1. Neural Network Model

The error between actual output values and target output values is calculated and propagated back toward hidden layer. The error is used to update the connection strengths between nodes, i.e. weight matrices between input-hidden layers and hidden-output layers are updated. The process is repeated till classification has to happen.

3 Feature Extraction

The four methods used for texture analysis and feature extraction:

- (1) Statistical methods based on the grey level co-occurrence matrix,
- (2) energy filters and edgeness factor,
- (3) Gabor filters, and
- (4) wavelet transform based methods.

Here will discuss co-occurrence matrix method.

3.1 Grey Level Co-occurrence Matrix (GLCM)

The elements of this matrix, $p(i,j)$, represent the relative frequency by which two pixels with grey levels "i" and "j", that are at a distance "d" in a given direction, are in the image or neighborhood. It is asymmetrical matrix, and its elements are expressed by

$$P(i,j) = \frac{P(i,j)}{\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} P(i,j)} \tag{1}$$

where Ng represents the total number of grey levels.

Using this matrix, Haralick (1973) proposed several statistical features representing texture properties, like *contrast*, *uniformity*, *mean*, *variance*, *inertia moments*, etc. Some of those features were calculated, selected and used in this study.

4 Image Classification Techniques

The conventional statistical approaches for image classification use only the gray values. Different advanced techniques in image classification like Artificial Neural Networks (ANN), Support Vector Machines (SVM), Fuzzy measures and Genetic Algorithms (GA) are developed for image classification. Here will discuss Artificial Neural Network technique.

4.1 Artificial Neural Network (ANN)

ANN is a parallel distributed processor [1] that has a natural tendency for storing experiential knowledge. Image classification using neural networks is done by texture feature extraction and then applying the back propagation algorithm.

5 Implementation Methodology

5.1 Feature Extraction

Texture and tone bear an inextricable relationship to one another. Tone and texture are always present in an image, although at times one property can dominate the other. For example, when a small area patch of an image has little variation of features of discrete gray tone, then tone is the dominant property. Important property of tone texture is the spatial pattern of resolution cells composing each discrete tonal feature. When there is no spatial pattern and the gray tone variation between features is wide, a fine textural image results. Texture is one of the most important defining characteristics of an image. It is characterized by the spatial distribution of gray levels in a neighborhood. In order to capture the spatial dependence of gray-level values, which contribute to the perception of texture, a two dimensional dependence, texture analysis matrix is considered. Since, texture shows its characteristics by both pixel and pixel values, there are many approaches used for texture classification. The gray-tone co-occurrence matrix is used for feature extraction. It is a two dimensional matrix of joint probabilities $P_d, r(i, j)$ between pairs of pixels, separated by a distance, d in a given direction $r1$.

- For finding textural features for every pixel in the image every pixel is considered as a centre and followed by a 3×3 window about that centre pixel.
- The gray-one matrix for that particular window is calculated and normalized.
- The gray level co-occurrence matrix namely, P_h, P_v, P_{rd} and P_{ld} for each pixel is then obtained. Here, P_h, P_v, P_{rd} and P_{ld} are respectively the $0^\circ, 90^\circ, 45^\circ$ and 135° nearest neighbors to a particular resolution cell.
- Standard deviation and mean are now obtained for each of these matrices and are used later to calculate the textural features.
- Now the particular entry in a normalized gray tone spatial dependence matrix is calculated for further reference, ie, $P(i, j), P_x(i), P_y(j), P_{x+y}(k)$ and $P_{x-y}(k)$.
- Using the formulas of the textural features, the angular second moment, contrast, correlation and variance are calculated.

The following equations define these features.

Notation

$P(i,j)$ (i,j) th entry in a normalized gray-tone spatial dependence matrix, $=P(i,j)/R$

$P_x(i)$ i th entry in the marginal-probability matrix obtained by summing the rows

$$\text{of } P(i,j), = \sum_{j=1}^{N_g} P(i,j)$$

N_g number of distinct gray levels in quantized image.

$$\sum_i \quad \sum_j \quad \sum_{i=0}^{N_g-1} \quad \sum_{j=0}^{N_g-1} \quad \text{respectively}$$

$$P_y(j) = \sum_{i=1}^{Ng-1} P(i,j) \tag{2}$$

$$P_{x+y}(k) = \sum_{\substack{i=1 \\ k=i+j}}^{Ng-1} \sum_{j=1}^{Ng-1} P(i,j) \quad k=2,3,4,\dots,2Ng \tag{3}$$

$$P_{x-y}(k) = \sum_{\substack{i=1 \\ k=i-j}}^{Ng-1} \sum_{j=1}^{Ng-1} P(i,j) \quad k=0,1,2,\dots,Ng-1 \tag{4}$$

Textural Features

1) Angular Second Moment:

$$F1 = \sum_i \sum_j \{P(i,j)\}^2 \tag{5}$$

2) Contrast:

$$F2 = \sum_{n=0}^{Ng-1} n^2 \left\{ \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} P(i,j) \right\} \tag{6}$$

3) Correlation:

$$F3 = \frac{\sum_i \sum_j (ij) P(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{7}$$

where μ_x, μ_y and σ_x, σ_y are mean and standard deviations of P_x and P_y .

4) Sum of squares - Variance:

$$F4 = \sum_i \sum_j (1-\mu)^2 P(i,j) \tag{8}$$

5.2 Training the Network Using Back Propagation Algorithm

Artificial neural network is trained using the Back propagation algorithm to classify the images.

The following assumes the sigmoid function $f(x)$

$$f(x) = \frac{1}{1 + e^{-x}}$$

The popular BKP algorithm is implemented using following steps:

Step 1: Initialize weights to small random values.

Step 2: Feed input vectors X_0, X_1, \dots, X_n through the network and compute the weighting sum coming into the unit and then apply the sigmoid function. Also, set all desired outputs d_0, d_1, \dots, d_n typically to zero except for that corresponding to the class the input is from.

Step 3: Calculate error term for each output unit as

$$\partial_j = y_j (1 - y_j) (d_j - y_j)$$

where d_j is the desired output of node j ; and y_j is the actual output.

Step 4: Calculate the error term of each of the hidden units as

$$\partial_j = x_j (1 - x_j) \sum_k \partial_k w_{jk}$$

where k is over all nodes in the layers above node j ; and j is an internal hidden node.

Step 5: Add the weight deltas to each of

$$W_y(t+1) = W_y(t) + \eta \partial_j x_i$$

All the steps excepting step 1 are repeated till the error is within reasonable limits and then the adjusted weights are stored for reference to the Recognition Algorithm.

6 Result

These four images are considered for training the neural network.



Image 1

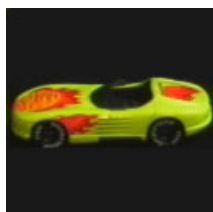


Image 2



Image 3

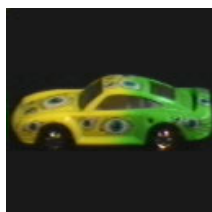


Image 4

This network tolerate error approximately equal to 0.0001. If any image is given as a query image to the trained network, it is classified into appropriate class.

If we give any image relevant to class I as a query image, first will extract textural feature of that image and give it as a input to trained neural network. A trained feed forward network then calculates the actual outputs and if it matches with the target output of a specific class then it retrieve that image from database.

For example, for rotated Image 1, the extracted textural features are

0.3777
0.9988
0.4964
0.0962

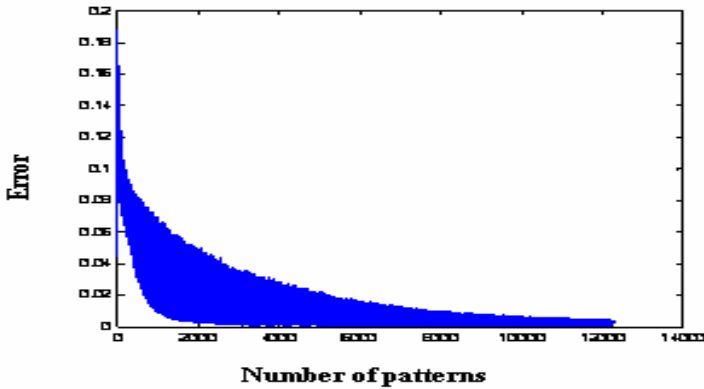
And actual outputs are

0.0003 = 0
0.0474 = 0

Hence the query image belongs to the Class I and retrieved from the database.

Table 1. Image database with extracted textural features as a input with desired output

Image	Input	Desired output	Class
1	0.3777 0.9988 0.4964 0.0962	00	I
2	0.4187 0.7646 0.3244 0.0713	01	II
3	0.2698 0.4999 0.6417 0.0898	10	III
4	0.3812 0.6555 0.3138 0.0638	11	IV

**Fig. 2.** Total Error during training**Fig. 3.** Query image retrieved from Class I

7 Conclusion

Artificial neural network is an efficient method for image classification. If neural network is defined properly and proper weights are assigned to the network then training is done accurately. The trained network gives better efficiency for image classification. Even if we rotate the image in any direction and give it as a query image to the trained neural network, it is classified to the appropriate class. The grey level co-occurrence matrix method is found to be an efficient method for textural feature extraction. For the better classification, we can add different features like color, shape with textural feature to train the Artificial Neural Network as well. We are also able to classify the images in more number of classes by modifying the definition of neural network and assigning different set of weights.

References

- [1] Zurada, J.M.: Introduction to Artificial Neural Networks System. Jaico Publishing House
- [2] Freeman: Artificial Neural Network Algorithm. Applications and Programming, Comp. and Neural Systems Series (1990)
- [3] Gonzalez, Woods: Digital Image Processing
- [4] Anderson, J.: An Introduction to Neural Network
- [5] Haykin, S.: Neural Network. a Comprehensive Foundation; a Computational Approach to Learning and Machine Intelligence. Macmillan, NY (1994)
- [6] Rao, V.V.: Artificial Neural Network. In: Concepts and Control Applications. IEEE Computer Society, Los Alamitos (1992)
- [7] Tian, B., Shaikh, M.A., Azimi-Sadajadi, M.R., Vonder Haar, T.H., Reinke, D.: A Study of Cloud Classification with Neural Networks using Spectral and Textural Features. IEEE Transactions on Neural Network 10 (January 1999)
- [8] Satellite Sensor Image Classification using Cascaded Architecture of Neuro Fuzzy Network. Geoscience and Remote Sensing, 1033 (January 2000)
- [9] Raghu, P.P., Yegnanarayan, B.: Supervised Texture Classification using PNN and Constraint Satisfaction Modes. IEEE Transactions on Neural Network 9, 516 (1998)
- [10] Segmentation of Color Textures, Pattern Analysis and Machine Intelligence, p. 142 (February 2000)
- [11] Haralick, Shanmugan: Textural Features for Image Classification. IEEE Transactions on Systems, Man and Cybernetics SMC 3(6), 610 (1973)

Data-Warehousing Applications in Manufacturing Industry – Applicable Solutions and Challenges Faced

Goparaju V. Ramesh¹, Sattiraju N. Rao², and Mogalla Shashi³

¹ IT Department, Visakhapatnam Steel Plant, Visakhapatnam-530 031, India

² ERP Department, Visakhapatnam Steel Plant, Visakhapatnam – 530 031, India

³ Department of CS & SE, College of Engineering, Andhra University, Visakhapatnam

gvr@vizagsteel.com, gvramesh67@gmail.com,

snrao@vizagsteel.com, smogalla2000@yahoo.com

Abstract. Data-warehousing is the computer application system that transforms a traditional intuitive decision making body into informed decision making organization. It provides key factors/facts across different chosen dimensions for ready consumption to the managers. The flexibility, provided by the data-warehouses for OLAP (On-line Analytical Processing) applications, offers a distinct edge to its practitioners over those who follow traditional way of browsing through conventional static reports for decision-making. However, the development and deployment of a proper data-warehousing solution in any industry is fraught with many challenges and more so in manufacturing industry in view of its complex processes and huge investments. Another important aspect is of identifying the right product to ensure that data-warehousing application can be built successfully for its effective use for on-line Analytical processing. This paper introduces the concept of data-warehousing, importance of it in a manufacturing industry, exploration of different types of solutions, and finally, the challenges faced in implementing the solution.

Keywords: Data-Warehousing, Business Intelligence, On-line Analytical Processing.

1 Introduction

The predominant processes involved in any manufacturing industry are metallurgical, chemical and mechanical in nature, like smelting, refining, rolling, wire drawing, forging, welding, chemical processing, distillation, etc. This processing domain is popularly known in IT (Information Technology) parlance as MRO (Maintenance Repair & Operation). The other divisions that work in close cooperation with MRO are

Materials Management – for handling input materials

Marketing & sales – for handling sales of products

Financials – for Financial Management

Human Resources – for Human resources management

As is well known, manufacturing industry involves huge investments in heavy machinery, equipment and employs thousands of manpower. In such a scenario, the decisions taken in these different business verticals will have far-reaching effects both on profitability and long-term strategic issues, like plant-expansion, manpower planning, etc. The industry should, therefore, perforce, transform to informed decision making rather than traditional intuitive decision making for effective results. This necessitates adequate, accessible, accurate information available at the doorstep / finger tip (literal meaning of finger at the mouse) of decision makers.

2 Data-Warehousing as a Tool for Informed Decision Making

The technology that makes the informed decision making possible is Data-warehousing, along with concomitant business intelligence tools for OLAP (On-line Analytical Processing). A data-warehouse is a repository of an organization's electronically stored data. It is a necessary prerequisite for making adequate and accurate information available & accessible to the managers. Every organization would certainly be having at least a rudimentary IT-infrastructure and OLTP IT-applications in the business domains mentioned above. The information contained in these IT-applications, different data-sheets generated in different offices using common office automation spread-sheets and small databases, and relevant data from the ubiquitous internet forms the basic ground-data for the data-warehousing applications.

The generic data-warehousing architecture consists of the following layers.

2.1 Operational/Business Database Layer

The source data for the data-warehouse — an organization's IT application systems, spreadsheets & data from Internet fall into this layer.

2.2 Data Access Layer

This layer acts as the interface between operational and information access layers. The extract, transform, load tools fall into this layer.

2.3 Metadata Layer

This layer consists of all the data about the sources of data, the paths leading to the target data, etc. This is equivalent to the data directories exist in databases. It is more detailed in data-warehouses when compared to the operational databases. There are dictionaries for the entire warehouse and sometimes dictionaries for the data that can be accessed by a particular reporting and analysis tool.

2.4 Information Access Layer

This layer consists of the data accessed for reporting and analyzing and the tools for reporting and analyzing data. Business intelligence tools fall into this layer.

3 Creation of Data-Warehouse

The creation of a data-warehouse includes following steps.

3.1 Understanding the Analytical Requirements of the Management

The IT personnel, who create the data-warehouse, should clearly understand the requirements of all the stake holders before venturing into creating the warehouse. They should look into the existing reports, the tactical and strategic requirements of the organization as a whole. The Software Requirement Specification should be prepared keeping all these in view. They should get the buy-in from the ultimate users for the document so prepared.

3.2 Data Understanding and Validation

Subsequent to understanding the requirements, the source data has to be identified, understood and validated over the frozen requirements. The source data can be anything from ERP (Enterprise Resources Planning) systems, spreadsheets, Internet, Taped back-ups, etc.

3.3 Data Extraction and Preparation

This is the most crucial step in that the data identified across different systems would generally be very disparate in nature, and needs lot of cleaning to be done for they contain lot of missing & invalid data. This step involves extraction (from source databases), transformation (Cleaning, Homogenizing, etc.).

3.4 Data-Warehouse Model Build

In this phase, the actual structure of data-warehouse is built. It involves creation of dimensions and facts. The dimensions provide the ways in which a fact/measure is analyzed. It generally indicates the master data, like supplier, time, region, etc. There are different hierarchies built in the dimensions. The fact table contains the actual numerical measures that need to be analyzed across different dimensions, like the sales value, purchase volume, cash disbursed, etc.

3.5 Data-Warehouse Deployment and Maintenance

This is the last step in the process of a data-warehousing solution. The cleaned and transformed data has to be loaded into the target data-warehouse. This process involves the scheduling and running of process flows for pushing data into the data-warehouses. The regular maintenance of the data-warehouse includes ensuring successful running of process flows, trapping errors, exceptions, overcoming the errors by going in for re-runs of scheduled process flows. Enhancing the dimensions and facts to take care of additional requirements, etc. also forms part of the maintenance of data-warehousing.

4 Typical Key Factors and Dimensions of Manufacturing Industry

The common facts/measures and related dimensions of interest to managers of different business domains in a Manufacturing Industry are tabulated in Table-1 below.

Table 1. Typical Facts / Measures and Dimensions across Different Business Domains in a Manufacturing Industry

Business Domain	Facts / Measures to explore	Dimensions
Maintenance, Repair & Operation	Operational Efficiency of Equipment, Maintenance History of Equipment, Productivity of different production departments, Product Quality parameters, Production parameters, etc	Product, Manufacturing section, Equipment type, time, etc.
Marketing & Sales	Fast moving products, Well run branches/regions, Valued customers, Net Sales Realization Market trends, etc.	Product, Region, Customer, Time, etc.
Materials	Vendor Classification, Fast moving spares, Economic order quantities, Estimates for future procurements, etc.	Supplier, Region, Material, Time, etc.
Human Resources	Employee Commitment, Per capita pay & perks, Employee Medical history, etc.	Employee Category, Employee Age, Employee Health Category, Time, etc.
Financials	Different Financial indices, Costing Sensitivity analysis, Cash Flow, Fund flow variances, etc.	Section, Account type, Schedule, Time, etc.

Quite often, the process of decision making requires data from different domains. For example, a marketing person would in all probability need information related to costing, which is the domain of operation and financials, so that he can optimize the net sales realization. Such requirements are known as cross-domain requirements. These also need to be considered while developing analytic requirements.

5 Finding Suitable Business Intelligence Solution

There are two components for a corporate BI (Business Intelligence) solution. They are a) Data-warehousing & b) OLAP reporting solution. The latter sits on top of the former. The OLAP solution provides necessary user interface for developing ready-to-create-and-use reports using data retrieved from the data-warehouses. The choices and the suitability of both the solutions are discussed below:

5.1 Data-Warehousing Solutions

There are two types of data-warehousing solutions available. They are Independent Data-Warehousing Solutions & ERP Based Data-Warehousing Solutions.

As the independent data-warehousing solution providers are not tied to any particular ERP solution, the capability of connecting to different types of applications be they legacy, ERP, office-automation, etc., is open and uniform. Whereas the ERP based solutions do, perforce, have better connectivity to their own ERP solutions.

In case of former, the logic for data extraction from the corporate databases for fact-related table data is entirely to be designed by the developers themselves. The reason being the basic corporate business application systems (Legacy systems) generally do not have the status-stubs (like changed data, new record, etc.) attached to each record for the sake of data extraction into data-warehouse. Therefore, it needs lot of coding both on part of source data systems and/or on part of ETL (Extraction, Transformation & Loading) systems for identifying the appropriate data for proper extraction into data-warehouse. Whereas, in case of the latter, as there is tight integration between basic ERP system and its data-warehousing system, the capability of finding the new/changed record is in-built and can be very simply implemented.

Based on this observation it can be concluded that the Independent data-warehousing solutions are best suited for the corporates, where the IT applications are legacy systems that grew over time across different platforms. Whereas ERP based solutions are best suited for the corporates, where the OLTP (On-line Transaction Processing) systems are predominantly ERP in nature.

5.2 OLAP Solutions

There are different types of OLAP solutions available. These include web-based reports, reports for the experts (Where intelligent users can generate reports on fly), for the lay-users (Where lay users can use already created OLAP reports), web-based portals showing the health of organization in stop-light and dashboard styles. The common features available in these solutions are as follows:

- Importing data stores from data-warehouses
- Creating tabular and matrix-type reports
- Flexibility of interchanging
 - the rows and columns
 - the order of rows and columns
- Filtering criteria
- Conditional formatting for exception reporting
- Sub-totaling / Totaling features
- Stop-Light & Dashboard facilities
- Mathematical & Statistical analysis
- Scheduling the report generation (This is a very good feature for the reports which take considerable amount of time for processing. These reports can be scheduled as per user requirements and can be viewed on fly)
- Capability of performing different operations, like

- Rolling up, drilling down,
- Dicey-chaining, slice-through, etc.

A judicious judgment can be made in selecting suitable solution for the OLAP reporting. The OLAP solutions are available as part of a data-warehousing solution, or as independent solution of its own. And, these can sit on top of number of different data-warehouses, created using different technologies (They contain suitable plug-ins to talk to different platforms). These can also be deployed successfully for OLAP applications.

6 The Challenges Faced in Development and Deployment of Data-Warehousing Applications in Manufacturing Industry

There are many challenges to face in implementing Data-Warehousing in any industry in general and manufacturing industry in particular. They are as follows.

Most of the manufacturing industries, especially steel, Non-Ferrous, etc. make products to inventory, but not to marketing order. The Sale orders do come much later to the process of manufacturing. This poses a great challenge in marrying the marketing orders to manufacturing data for effective analysis of marketing issues with respect to manufacturing dimensions.

This challenge can be overcome by following proper identification of product both in in-process inventory stage and final production stage so that unambiguous linkage can be established between production order and marketing sale order, both in order generation stage and logistics phase.

In most of the manufacturing industries, direct raw materials form bulk of the cost of production. And, data related to these raw materials is very crucial in taking important decisions related to logistics, cost of production, inventory management, etc.. However, these basic raw materials for many of the manufacturing Industries are bulk handled materials, and the weight and quality related data (Chemical composition) obtained directly from measuring scales are quite inaccurate. This necessitates lot of approximation in integrating this data, thus posing a major challenge for data-warehouse builders. The factoring of approximation is also not uniform, and is seasonal in nature (For example, the monsoon season is very difficult season for maintaining the instrumentation attached to bulk handled raw materials on conveyors, the spontaneous combustion of burning that takes place during summers in coal yards results in lot of disparities between laboratory data and actual quality of coal).

This challenge can be overcome by going in for state-of-the-art weighing scales coupled with strict maintenance and calibration regimen for getting accurate results.

The total equipment/assemblies in many manufacturing industries are very large in number. It is very time-consuming and difficult to track their maintenance schedules / repair cycles and spares planning, an important aspect for maintenance, repair and operation domain.

This challenge can be overcome by careful planning of all the equipment, assemblies and their spares and consumables right to the lowest detail, so that the linkage / hierarchies can be built in data-warehouse for effective roll-up and drill-down operations.

In addition to integrated data provided through ERP environment, some of the business activities do still take place in the stand-alone PCs / Workstations using office automation systems, like Spreadsheets, Local databases, etc. All this data is quite obviously very ad-hoc, un-parameterized and unstructured in nature. The assimilation of such data poses a great challenge for any creator of data-warehouse.

This challenge can be overcome by going in for fixed templates of spreadsheets avoiding sub-totals, merger of multiple columns, etc. in the basic data related spreadsheets (Macros can be written in other sheets for presentation of data in the required format for the spread-sheet users.). This procedure can make the import of raw-data from spreadsheets into the data-warehouse in a proper manner.

Most of the IT application systems in manufacturing industry evolve over different time horizons. This results in creation of new attributes to the existing entities over a period time. The values of these attributes would be empty (null) for the past tuples. A judicious call has to be taken for replacing these missing values with averaged out values without impairing the outcome.

The other challenge is that related to aggregation operations performed in rolling-up operations. In most of the cases, the requirements for aggregation along a dimension vary from one dimension to the other. Most of the Data-warehousing softwares do not provide the feature of selective aggregation. This issue is illustrated in an example here under.

6.1 Example of Selective Aggregation Problem in Data-Warehousing

Fact Table – Account wise transaction amount & Balance

Dimension Tables – Account dimension & Time dimension.

Table 2. Account Dimension Table

Group Code	Group description	Account Code	Account Description
522	Advances	522701	Advances to Suppliers
522	Advances	522702	Advances to Employees

Table 3. Account-wise Transaction Amount & Balance Fact Table

Month	Account code	Transaction-Amount, Rs.	Account Balance, Rs.
Jan-2009	522701	10000	15000
Feb-2009	522701	25000	40000
Mar-2009	522701	-12000	28000
Jan-2009	522702	15000	32000
Feb-2009	522702	-25000	7000
Mar-2009	522702	15000	22000

The account dimension is having a hierarchical arrangement as follows:

- Account Code-Group Code
- Time Dimension - Date-Month-Quarter-Year

The business rules for aggregation are as follows:

- The transaction-amount can be aggregated in any hierarchy
- The Account balance amount can be aggregated along account dimension, but not time dimension.

When this data is put in data-warehousing and aggregation (Sum) is applied, the aggregated results look as shown below:

Table 4. Aggregation along Time-Dimension

AGGREGATION ALONG TIME DIMENSION		
Account description	Transaction Amount, Rs.	Account Balance, Rs.
522701	23000	83000
522702	5000	61000

These aggregated account balances are in actuality wrong as account balances cannot be aggregated along time dimension (These should be the latest account balances, i.e., 28000, 22000 respectively). But, as this check cannot be put in most of the available softwares, this wrongful aggregation cannot be prevented in the automatically generated OLAP reports.

Table 5. Aggregation along Account Dimension

Month	Transaction Amount, Rs.	Account Balance, Rs.
Jan-2009	25000	47000
Feb-2009	0	47000
Mar-2009	3000	50000

! – These aggregations are correct, as the account balances can be aggregated along the account codes.

These aggregated account balances are correct, as the balance can be aggregated along the account codes.

Some of the data-warehousing have now come up with the concept of non-cumulative key-figures. These may facilitate overcoming these issues in a better manner.

One of the important challenges to overcome is the aspect of different aggregation techniques to be applied over different facts of same fact-table. For example, a fact table contains item-id, quantity and unit rate as facts/key-figures. The aggregation statistic of sum may convey some meaning for quantity (provided the units are same for the items under consideration), but the same aggregation statistic may not convey proper meaning for unit-rate as, Sum of unit-rate is not useful. Rather, weighted average is a much better

option for unit-rate. Such a requirement of different aggregation statistic functions for different facts of same fact-table is not available in conventional OLAP-tools.

In some of the operations of core-manufacturing areas, like in Steel melting shop, the number of dimensions, like tapping time, tapping temperature, different constituents of chemical composition, tundish related parameters, oxygen blowing time, temperatures, etc., are so large that it takes lot of processing time and processing power to warehouse all that data. This necessitates deployment of data-warehousing applications in servers having higher processing power.

The sorting sequence is another – albeit trivial – challenge to overcome in some cases. For example, sorting an OLAP report based simply on an item number or item name may not be useful. Whereas, sorting based on type of material may be more useful. However, in case the type of material is not an attribute in the report, then this sorting order would be missed.

Some of the data-warehousing application provide a concept called Alternate-Sort order, and based on which this problem can be solved to some extent.

The process of data extraction process flows, conversion and scheduling the process flows is a real challenge for any data-warehousing application. It calls for real application of proper skill and practical knowledge in developing, testing and implementing these routines. The monitoring of these flows as per predetermined schedules is a real task to be performed everyday by the administrators in order for proper data extraction, transformation and loading.

There is a major challenge, not technical in nature, but more of a socio-psychological, inter-personal and managerial in nature. The basic lethargy, clinging to status quo, resisting change, Fear of adapting to new, modern and effective decision making tools on the part of managers is a major hurdle in implementing any new system. And, data-warehousing is no exception to it. The very process of logging into the system, navigating across business analytical pages may not be liked and not be practiced by hard-to-change-managers. It is the responsibility of the BI developers to conduct series of awareness programs, user training sessions and workshops to drive home the advantage of using these systems for their own benefit.

These are the challenges that are to be addressed for successful creation and deployment of data-warehousing system in a manufacturing industry.

7 Conclusion

The importance of data-warehousing has become vital in view of the increased necessity of informed decision making. Today, business managers need to act on dynamic data analysis encompassing different data domains rather than relying on static reports of individual domains. However for effective decision making, one needs to understand the challenges to face and evaluate different options available. Hence, choice of proper data-warehousing solution and its successful implementation play an important role for building business intelligence required for effective decision making in an organization.

References

1. Ponniah, P.: Data Warehousing Fundamentals. John Wiley & Sons, Inc., Chichester (2001)
2. Inmon, W.H.: Building the Data Warehouse. John Wiley & Sons, Inc., Chichester (2002)

Application of Clustering in Virtual Stock Market

Kavita M. Gawande¹ and Sangita C. Patil²

¹ K.J. Somaiya Institute of Engineering,
& Information Technology,
Sion, Mumbai
+919892829102
kavita.bathe@gmail.com

² K.J. Somaiya Institute of Engineering
& Information Technology,
Sion, Mumbai
+919769298339
sangita_nemade@rediffmail.com

Abstract. The groups of companies are categorized due to the splitting of financial markets into diverse sectors. The market scenario much is directory related to the rates of share prices that are expected to move up or down. There are several factors that influence the individual share prices.

A profitable share market always witnesses the trend of the rates plunging down. Within the market sector, there is anticipation for the shares to move for the most part, mutually. In this case, our aim is to identify the groups of shares that do move collectively and identifiable in conditions of corporate activity. A hierarchical clustering algorithm, tree GNG, have identified groups of companies that cluster in clearly identifiable sectors. This categorization of sector scheme is accepted all over the world.

Keywords: Data mining; clustering algorithm.

1 Introduction

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

There has been recent interest in the discovery of structure in stock market from an analysis in time series of closing prices of shares traded within the market. The methods of discovering structure are generally based on the clustering algorithms which may be useful in producing a theoretical description of financial markets and explaining the economic factors that affects specific groups of stocks.

Our approach uses a neural inspired clustering algorithm to learn a topology preserving the mapping of data while maintaining the history of learning process which gives hierarchical structure. We use the Euclidean norm combined with normalized data for the measurement of distance. Thus, the application of our hierarchical clustering method, combined with data normalization of financial time series data, is the contribution of this paper.

Data clustering has been used for the following three main purposes:

- *Underlying structure*: to gain insight into data, generate hypotheses, detect anomalies, and identify salient features
- *Natural classification*: to identify the degree of similarity among forms or organisms.
- *Compression*: as a method for organizing the data and summarizing it through cluster prototypes.

2 Review of Clustering Algorithms

2.1 K – Means Algorithm

Let $\{X_i = x, i = 1, \dots, n\}$ be the set of n d -dimensional points to be clustered into a set of K clusters, $\{1, \dots, k\}$ $C_k = c$ $k = K$. K-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. Let m be the mean of cluster ck . The squared error between m and the points in cluster ck is defined as

$$J(C_k) = \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \quad (1)$$

The goal of K-means is to minimize the sum of the squared error over all K clusters,

$$J(C) = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \quad (2)$$

Minimizing this objective function is known to be an NP-hard problem (even for $K = 2$). Thus K-means, which is a greedy algorithm, can only converge to a local minimum, even though recent study has shown with a large probability K-means could converge to the global optimum when clusters are well separated. K-means starts with an initial partition with K clusters and assign patterns to clusters so as to reduce the squared error. Since the squared error always decrease with an increase in the number of clusters K (with $J(C) = 0$ when $K = n$), it can be minimized only for a fixed number of clusters. The main steps of K-means algorithm are as follows:-

1. Select an initial partition with K clusters; repeat steps 2 and 3 until cluster membership stabilizes.
2. Generate a new partition by assigning each pattern to its closest cluster center.
3. Compute new cluster centers.

2.2 Growing Neural Gas Clustering Algorithm

Growing Neural Gas (GNG) algorithm is a inspired clustering algorithm that produces clustering schemes with no hierarchical structure. GNG incrementally adapts, grows and prunes a graph consisting of nodes and edges. The nodes of the graph are of the same dimensionality (D) as the input space. The topological structure are defined as the nodes that are adjacent in the graph.

The Growing Neural Gas (GNG) is the method where number of neurons is not immutable input parameter but is changed during the competition. Connections between neurons are not permanent as well. The result of competition could be then set of separate neural networks covering some region of the input data.

In the beginning the network itself contains only two neurons a_1 and a_2 representing two randomly chosen input patterns. Denote set of neurons as A and set of connections as C which is empty in the beginning.

Competition: The pattern x is presented to the network. The winner s_1 of competition and the second nearest s_2 neurons are determined using equation

$$C = \arg \min_{a \in A} \{ \|x - w_a\| \} \quad (3)$$

If there was not a connection between neurons s_1 and s_2 then it is created ($C = C \cup \{(s_1, s_2)\}$). The age of the connection is set or updated to 0 ($\text{age}(s_1, s_2) = 0$). The squared distance between the winner and the pattern is added to local error variable. $E_{s_1} = \|x - w_{s_1}\|^2$.

Adaptation: The weight vectors of the winner and its direct topological neighbours N_{s_1} are adapted by fractions β and η of the distance to the input pattern. This is analogous to the Kohonen's rule.

$$\begin{aligned} \Delta w_{s_1} &= \beta (x - w_{s_1}) \\ \Delta w_i &= \eta (x - w_i) \quad \forall i \in N_{s_1} \end{aligned} \quad (4)$$

The age of all connections leading from the winner neuron are increased by 1

$$(\text{age}(s_1, i) = \text{age}(s_1, i) + 1 \text{ for all } i \in N_{s_1}). \quad (5)$$

Removing: If there exist some connections with age greater than given a \max then all are removed. If this step results in neurons with no connections then remove also these standalone neurons.

1. First of all, the neuron p with the largest accumulated local error is determined using following equation.

$$p = \arg \max \{ E_a \} \quad (6)$$

Among the neighbours of neuron p determine neuron r with largest accumulated local error.

Inserting new neurons: If the number of processed patterns reached an integer multiple of given parameter λ then new neuron is inserted using following steps:

1. First of all, the neuron p with the largest accumulated local error is determined using following equation.

$$p = \arg \max_{a \in A} \{E_a\} \quad (7)$$

Among the neighbours of neuron p determine neuron r with largest accumulated local error.

2. Insert new neuron q to the network ($A = A \cup \{q\}$) and set its weight to the mean value of p and r weight vectors.

$$w_q = \frac{1}{2} (w_p + w_r) \quad (8)$$

3. 1) Insert new connection between new neuron q and neurons p and r

$$(C = C \cup \{(p, q), (r, q)\}) \quad (9)$$

2) Remove the old connection between neurons p and r

$$(C = C - \{(p, r)\}). \quad (10)$$

4. Local accumulated error variables of neurons p and r are decreased by given fraction α .

$$\Delta E_p = -\alpha E_p \quad \Delta E_r = -\alpha E_r \quad (11)$$

The accumulated error variable of the new neuron q is set to the mean value of neurons p and r accumulated error variables.

5. Local accumulated error variables of all neurons in the network are decreased by given fraction β .

$$\Delta E_a = -\beta E_a \quad \forall a \in A \quad (12)$$

These several steps proceed until pre-defined termination condition is met.

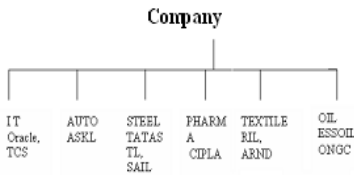
2.3 Advantage of GNG

- On the other side, GNG can either add nodes where the local error is too large or remove nodes that have no relation to the rest (no existing connection).
- GNG adds nodes after the pre-defined chunk of steps passes which saves time.
- GNG proceeds with an extended goal.
- The GNG algorithm is covering all patterns with the nodes and moreover it covers the cluster's topology with the edges.

3 Stock Market Overview

3.1 Stock Market Structure

Financial markets are complex systems. To aid the understanding of the core business activities of the companies within these complex systems, the constituent companies

Table 1. Classification Scheme**Fig. 1.** Global Classification System**Fig. 2.** Company categories

Company name	Script ID	Sector
Reliance Industries	RIL	TEXTILE
Oil and natural gas limited	ONGC	OIL
Tata Steel	TATASTL	STEEL
Tata Consultancy services	TCS	IT
Larsen and Turbo	LNT	MACHINERY
Steel Authority of India Ltd.	SAIL	STEEL
Asian Brown Braveries	ABB	MACHINERY
ACC	ACC	CEMENT
Ashok Leyland	ASKL	AUTO
Essair Oil	ESSOIL	OIL
Cipla	CIPLA	PHARMA
Oracle	Oracle	IT
Arvind Mill	ARND	TEXTILE

within a stock market are assigned to one-of-many possible sectors. One universally accepted classification scheme is FTSE Global Classification System which is based on the division of a market into Economic Groups, Industrial Sectors and Industrial Sub-sectors. The aim of the classification is to allocate a company to the sub-sector that most closely describes the nature of its business.

3.2 Stock Market Data

The closing price for the constituent companies of the FTSE are collected for the total period (T). We denote $P_{i(t)}$ as the closing price of company i , ($i= 1, \dots, N$) where

N =total no. of companies considered on day t , ($t=1, \dots, T$), and the logarithmic return, S_i , on the share price after time interval Δt (where Δt is one trading day) as

$$S_{i(t)} = \ln(P_{i(dt)}) - \ln(P_{i(t)}) \tag{13}$$

For the stock market data examined in this paper, it is possible to normalise the data either i) daily with the division by the range of S for all the companies at time t , or ii) normalise the logarithmic return $S_{i(t)}$ against the range of S_i in T . The results presented in this paper are based on the logarithmic return of N . $S_{i(t)}$ normalised by the range of S_i , with a Euclidean metric.

4 Experimental Results

We have applied two algorithms, K-means and GNG, on data to view their ability to detect clusters in unknown data. The software prepares chart/graph of the trading & transactions that are implemented by clients to analyze future price movement of the stock. A client can analyze movement of a stock based on groups/clusters formed. The analysis done is that if a price of overall cluster/group increases, then the price of stocks in that cluster/group increases or remains the same but will not decrease. Similarly analysis is done for decreasing prices in the cluster/group. We have applied two algorithms, K-means and GNG, on data to view their ability to detect clusters in unknown data. Figure 3 and Figure 4 depicts the results.

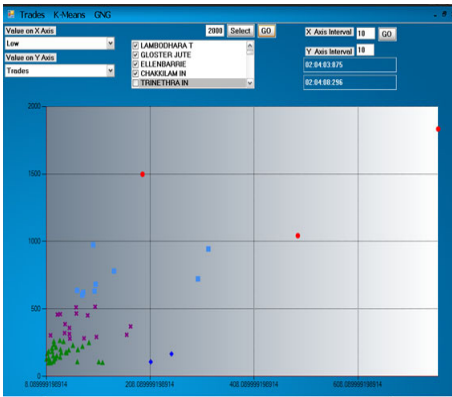


Fig. 3. Clusters using K means

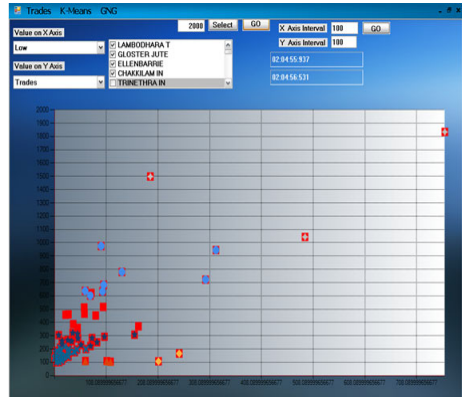


Fig. 4. Clusters using GNG

K means: Here the number of clusters are predefined & number of companies can be variable. Companies are added first and then clustering algorithm as a whole is applied. Not too much memory space is required. Less parameters are required. As number of clusters are predefined it is static. A graph showing various clusters of different companies satisfying a certain criteria is displayed using k means algorithm. Values on X and Y axis could be any one of low, high, trades, shares, turnover, open and close. A client can analyze data of n number of companies ($n < 3070$, since database has only 3070 scripts). X and Y axis intervals can be adjusted for better view of the graph. Always 5 clusters are formed as K-MEANS is static (number of Clusters are defined to be 5).

GNG: Here user has to give no of companies (variable) & depending on their proximity the clusters (variable) are formed. Here companies are added dynamically i.e. during execution/in run time of the algorithm. Lots of parameters are required. It takes greater memory space since it is dynamic.. A graph showing various clusters of different companies satisfying a certain criteria is displayed using GNG algorithm. Values on X and Y axis could be any one of low, high, trades, shares, turnover, open and close. A client can analyze data of n no of companies ($n < 3070$, since database has only 3070 scrips). X and Y axis intervals can be adjusted for the better view of the graph. No. of Clusters varies according to the input. As it is dynamic no. of clusters are not specified. No. of clusters are decided during run time. GNG algorithm when applied on data forms 7 clusters. Time required to form clusters is 594ms.

4.1 Comparison of GNG and K-MEANS Algorithms

From figure 5 ,we can analyze that GNG is better than K-MEANS as it takes less time to execute and is more accurate since more number of clusters are formed. The result obtained is very much useful in stock market.

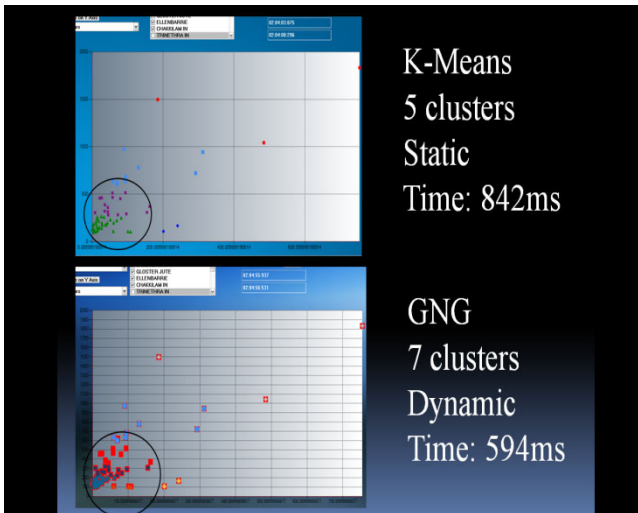


Fig. 5. Comparison of K-means & GNG

5 Cryptography Used

5.1 RSA Cryptography

Cryptography is the science of keeping data secure. RSA is an encryption algorithm used for data security. Here we have implemented it on buying/selling quantity/rates. Encryption is done for storing transactions in database & decryption is done during fetching from database.

ID	Symbol	Date	Price	Volume
1	I	20100409	00:19:13	1
2	I	20100409	00:23:26	1
3	I	20100409	00:50:34	0
4	I	20100409	00:51:49	1
5	I	20100409	00:52:10	1
6	I	20100409	00:53:18	0
7	I	20100409	00:53:20	0
8	I	20100409	00:54:12	1
9	I	20100409	13:37:39	1
10	I	20100409	00:58:15	0
11	I	20100409	23:22:26	1
12	I	20100409	23:26:50	1
13	I	20100409	23:26:50	1
14	I	20100421	16:19:41	0
15	I	20100421	16:19:41	1
16	I	20100421	16:19:51	0
17	I	20100421	16:19:52	0
18	I	20100421	16:19:52	1
19	I	20100421	16:20:37	0
20	I	20100421	16:20:37	1
21	I	20100421	16:41:11	0
22	I	20100421	16:41:11	1
23	I	20100421	16:41:20	0
24	I	20100421	16:41:20	1
25	I	20100421	20:31:44	0
26	I	20100421	20:31:44	1

Fig. 6. Results of Cryptography

6 Conclusion

Predication of stock market prices can be useful in many context. Most stocks tend to increase in price, and when the market goes down, they tend to decrease in price. From the comparisons between different clustering methods such as K-Means and GNG, we see that utilizing GNG algorithm is good idea in the clustering domain. The results obtained using this method shows that neural networks are able to give more accurate results. They help to preserve the topology which improves the results.

References

1. Mantegna, R.N.: Hierarchical structure in financial markets. Eur. Phys. J. B 11, 193–197 (1999)
2. Bonanno, G., Caldarelli, G., Lillo, F., Mantegna, R.: Topology of correlation-based minimum spanning trees in real and model markets. Rev. E 68, 046130 (2003)
3. Fritzke, B.: A growing neural gas network learns topologies. Neural Information Processing Systems, 625–632 (1995)
4. Doherty, K.A.J., Adams, R.G., Davey, N., Pensuon, W.: Hierarchical Topological Clustering Learns Stock Market Sectors (2005)
5. Hertz, J., Krough, A., Palmer, R.G.: Introduction to the theory ofneural computation. Addison-Wesley Publishing Co., Reading (1991)
6. Heinke, D., Hamker, F.H.: Comparing neural networks: a benchmark Growing Neural Gas. Growing Cell Structures and Fuzzy ARTMAP
7. Doherty, K.A.J., Adams, R.G., Davey, N.: Hierarchical Growing Neural, Coimbra, pp. 140–143 (2005)

8. FTSE, Ground rules for the Management of the FTSE Global the finer grained 100-node network. In addition to extracting Classification System,
http://www.ftse.com/indices_marketdata/groundrules/global-classification-ground-rules.pdf
(accessed June 20, 2005)
9. FTSE, UK Index Series,
http://www.ftse.com/indices_marketdata/uk_series/index_home.jsp#ftse
10. <http://www.clusty.org>

Distribution of Loads and Setting of Distribution Sub Station Using Clustering Technique

Shabbiruddin and Chakravorty Sandeep

Department of Electrical & Electronics Engg., Sikkim Manipal Institute of Technology, Sikkim
shabbiruddin85@yahoo.com
sandeep_chakravorty@yahoo.com

Abstract. Choosing an optimum location of a distribution substation and grouping the various load points to be fed from a particular distribution substation has always been a concern to the distribution planners. A lot of work has been carried out in this regards but all have made either the use of man machine interface or have made some approximations. Here in this paper we present a Hard Clustering method for grouping the various load points as per the number of distribution transformers available. The method further gives an optimum location of the distribution substation taking into aspects the distances of the various load points that it is feeding. The results of the discussed techniques will lead to a configuration of substations that will minimize substation construction cost. It will further lower long range distribution expenses as it will lead to optimum feeder path.

Keywords: Distribution planning, Hard Clustering.

1 Introduction

In general, the decisions in the planning of power distribution system include:

- Optimal location of substations
- Optimal allocation of load
- Optimal allocation of substation capacity

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. Unfortunately only near optimal solutions have been obtained by these mathematical programming methods because almost every method has made some approximations on the model of distribution planning, moreover 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 substation to meet the load demands of 13 load points whose coordinates and MVA demands are given. Similar work has been carried out by Belgin Turkay and Taylan Artac [1], work has also been carried out by J.F.Gomez *et.al.*, [2]. In all the above cases planning of laying the feeders or distribution planning has been done either by man machine interface or heuristic algorithm.

Here in this paper we suggest the location of the substation and the various load points to be fed by the substation by means of hard clustering technique. No man machine interface is required for determining the clustering of loads to be fed by a substation as indicated in the previous works.

A complete survey of the proposed techniques for the solution of the planning problem of primary distribution circuits can be found in [8] and [9]. Initially 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.

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 substation 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 involves the task of dividing data points into homogeneous classes or clusters so that items in the same class are as similar as possible and items in different classes are as dissimilar as possible. Clustering can also be thought of as a form of data compression, where a large number of samples are converted into a small number of representative prototypes or clusters. Depending on the data and the application, different types of similarity measures may be used to identify classes, where the similarity measure controls how the clusters are formed.

In this study Hard clustering method is used to divide various load points into clusters. A substation is placed for each of the classes obtained from the clustering. Hard Clustering Method (HCM) is a data clustering technique wherein each data point is divided into distinct clusters, such that each data element belongs to exactly one cluster. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters.

2 Proposed Methodology

Architecture of Hard Clustering Method(HCM).

The HCM algorithm attempts to partition a finite collection of n elements $X = \{x_1, \dots, x_n\}$ into a collection of c hard clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of c cluster centres $C = \{c_1, \dots, c_c\}$

and a partition matrix , where each element u_{ij} tells the degree to which element x_i belongs to cluster c_j . Like the k-means algorithm, the HCM aims to minimize an objective function.

- Number of classes are considered depending on how many cluster is needed to be obtained from the given data points.
- The cluster centers are obtained by.” $V_{ij}=(U_{ij}*(X_{ij} \text{ or } Y_{ij}))/(\text{no. of elements for that particular set})$ ”
- The distance is found out from sample to the centre,

$$D_{ij}=\text{sqrt}((X_{ij}-V_{ij})+(Y_{ij}-V_{ij}))$$
- The minimum of the two distance is found out and feeded into the matrix.
- The iteration is continued unless the last two iteration produces the same result.
- The last cluster center obtained is the appropriate location.

Case Study:

Let us have the problem discussed by S. Chakravorty *et.al.*,[15] where a thirteen load points are to be fed from two substations depending on the capacity and the load demands. The table below shows the data of the thirteen load points considered.

Table 1. Showing 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

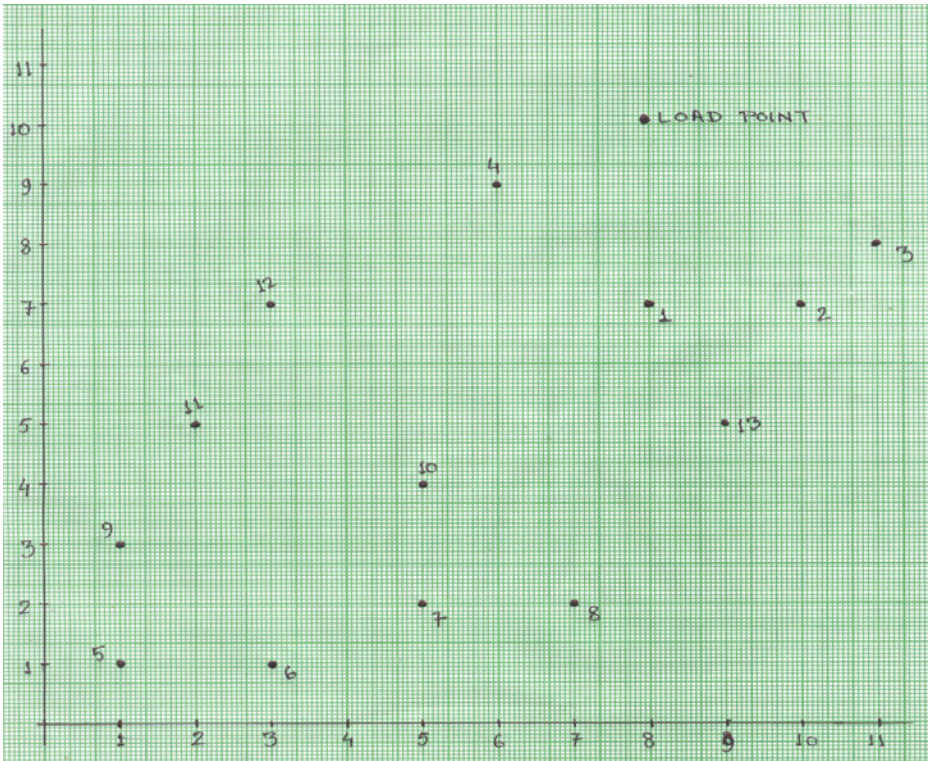


Fig. 1. Pictorial representation of the problem

We intended to feed the given thirteen load points with the help of two substations. So we divide these load points into two clusters so that a substation could be placed for each of the cluster obtained.

3 Result

The 4th and 5th iteration are giving the same results. Thus from the results it is clear that load points (1,2,3,4,13) are in class C1 while load points(5,6,7,8,9,10,11,12) are in class C2.

The cluster center obtained are (8.8, 7.2) and (3.375, 3.125). These cluster centers could be used for placing the substation as they are optimum from the point of view of distance.

Table 2. Showing the initial assumption

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13
C1	1	1	1	1	0	0	1	1	0	1	1	1	1
C2	0	0	0	0	1	1	0	0	1	0	0	0	0

Cluster center obtained (6.6, 5.6), (1.67, 1.67).

Table 3. Showing the result of Iteration 1

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13
C1	1	1	1	1	0	0	0	1	0	1	0	1	1
C2	0	0	0	0	1	1	1	0	1	0	1	0	0

Cluster center obtained (7.375, 6.125), (2.4, 2.4).

Table 4. Showing the result of Iteration 2

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13
C1	1	1	1	1	0	0	0	1	0	0	0	1	1
C2	0	0	0	0	1	1	1	0	1	1	1	0	0

Cluster center obtained (7.71, 6.42), (2.833, 2.67).

Table 5. Showing the result of Iteration 3

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13
C1	1	1	1	1	0	0	0	0	0	0	0	0	1
C2	0	0	0	0	1	1	1	1	1	1	1	1	0

Cluster center obtained (8.8, 7.2), (3.375, 3.125).

Table 6. Showing the result of Iteration 4

	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13
C1	1	1	1	1	0	0	0	0	0	0	0	0	1
C2	0	0	0	0	1	1	1	1	1	1	1	1	0

Cluster center obtained (8.8, 7.2), (3.375, 3.125).

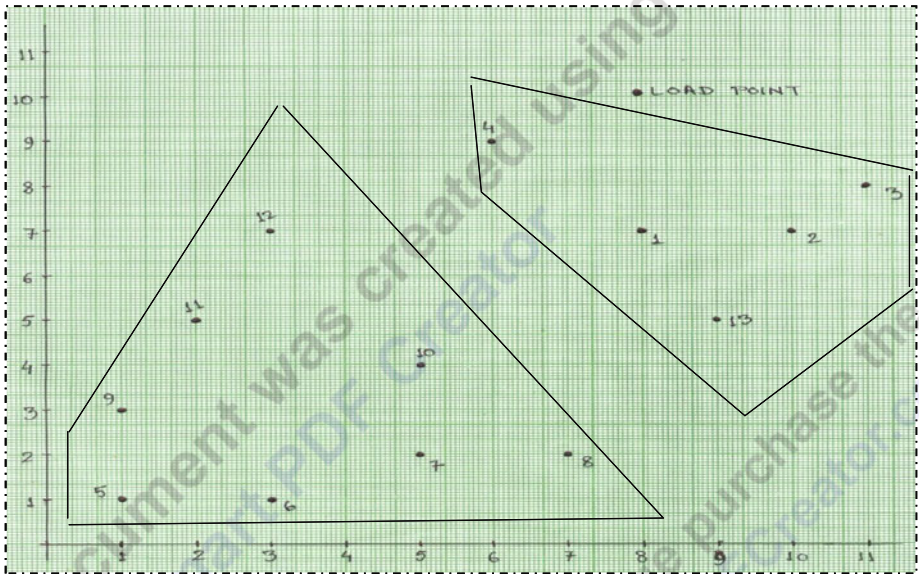


Fig. 2. Showing the result in pictorial form

4 Discussion and Conclusion

In the paper of S.Chakravorty *et.al.*,[15] the load distribution was proposed using the concept of genetic algorithm in which the capacity of the substation 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, the clustering of the load points is done irrespective of the capacity of the substation. One may decide on the capacity of the substation depending on the load points required to be fed from the substation.

A new methodology, based upon the HCM algorithm, is proposed for the planning of electrical power distribution system. Thus by applying Hard Clustering method, various load points which are at different location can be grouped into number of clusters depending on the number of distribution substations available. Also the location of the substation can be determined. The technique suggested is simpler than all the existing methods. 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. Turkey, 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 Substations, 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 Substation 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 Electrica, Ph.D. dissertation, 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(2), 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. Dhar, S., Ray, A., Bera, R., Sur, S.N., Ghosh, D.: A Complete Simulation Of Intra Vehicle Link Through Best Possible Wireless Network. *International Journal of Computer and Electrical Engineering* 2(4), 673–681 (2010)
24. Ray, A., et al.: Process Cost Prediction: A Soft Computing Approach. *International Journal of Intelligent Computing and Cybernetics* 3(3), 431–448 (2010)

Landscape of Web Search Results Clustering Algorithms

Ujwala Bharambe¹ and Archana Kale²

¹ Thadomal Shahani Engineering College, PG Kher Marg, TPS III,
Bandra (W), Mumbai -50
ujwala.b@gmail.com

² Thadomal Shahani Engineering College, PG Kher Marg, TPS III,
Bandra (W), Mumbai -50
archiekk@yahoo.co.in

Abstract. Searching for information on the Web has attracted great attention in many research communities. Due to the enormous size of the Web and low precision of user queries, results returned from present web search engines can reach hundreds or even hundreds of thousands documents. Therefore, finding the right information can be difficult if not impossible. One approach that tries to solve this problem is by using clustering techniques for grouping similar document together in order to facilitate presentation of results in more compact form and enable thematic browsing of the results set. Web Search Results clustering is about efficient identification of meaningful, thematic groups of documents in a search result and their concise presentation. This paper is an introduction to the problem of web search results clustering and we have a brief survey of previous work on web search results clustering and existing commercial search engines using this technique, and propose the possibility of future research direction.

Keywords: Web mining, Search engine, Clustering, Information retrieval, Web Search.

1 Introduction

With its explosive growth, the World Wide Web (the Web) has become an immense resource of textual data, images and other multimedia content. For efficient access and exploration of useful information, appropriate interfaces to search and navigation through this enormous collection are of critical need.

An alternative and increasingly popular method is also search results clustering [2] [23]. Search results clustering is a process of organizing document references returned by a search engine into a number of meaningful thematic categories. In this setting, in response to a query "Mumbai", for example, the user would be presented with search results divided into such topical groups as "University of Mumbai", "Mumbai city" "Mumbai travel", "Mumbai Map", etc. Users who look for information on a particular subject will be able to identify the documents of interest much quicker, while those who need a general overview of all related topics will get a concise summary of each of them.

The purpose of clustering is not to improve precision in search results, but rather an attempt to make search engine results easier to browse. Search results clustering is an attempt to apply the idea of clustering to document references (snippets) returned by a search engine in response to a query. Thus, it can be perceived as a way of organizing the snippets into a set of meaningful thematic groups. There are several ways in which end users can benefit from such a clustered view: Zamir [1] has identified several key requirements for web document clustering method. 1. Fast access to relevant documents 2. Broader view of the search results 3. Relevance feedback functionality 4. Relevance 5. Browseable Summaries 6. Overlapping clusters. 7. Snippet tolerance 8. Speed 9. Incremental processing.

Having understood the problem, we may define it more formally: Let there be a certain number N of *search results* as returned by a traditional search engine in response to some *query*. Each search result represents a document in the internet and is composed of the document's URL, an optional title and short text relevant to the document's contents, called a *snippet*. We may assume that the search engine's algorithms work in such way, that snippets are descriptive about the topic of the documents they represent, even if they are not part of those documents' body. For some documents snippets may not exist at all, or be short and form no valid language constructs (sentences). Let there be P possibly overlapping *topics* in the set of documents in the result. One document may belong to more than one topic. Topics are *semantically groups*, i.e. they represent the real *meaning* of the documents, not necessarily statistical distribution of terms (although this may be true). The set of topics may contain relations - in particular, hierarchical dependencies may exist.

2 Landscape of Clustering Algorithm

2.1 Clustering Approaches (Traditional Approach)

The classic approach attempts to adopt the well-known clustering algorithms, originally designed for numerical data, such as Hierarchical Agglomerative Clustering (HAC) or K- means, to the data of textual type. The algorithms require that for every two objects in the input collection, a similarity measure be defined. The measure, which is usually calculated as a single numerical value, represents the "distance" between these objects. Objects that are "close" to each other in this sense will be placed in the same cluster.

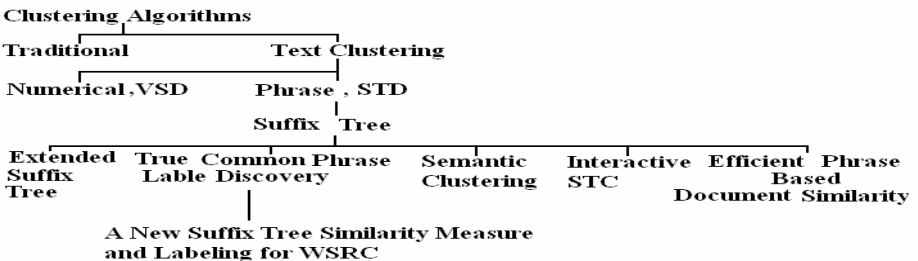


Fig. 1. Classification of various clustering algorithm

The main distinction between web search result clustering (WSRC) and classical document clustering is in assessment of algorithm's quality. Classical methods are usually evaluated based on mathematical proofs of correctness plus recall and coverage on test data. SRC is a fully user-oriented field and is evaluated by *subjective* assessment of the quality of produced clusters. More formal methods should be elaborated with time, but users' judgment about usefulness of produced clusters should always be the most important voice. In other words, any algorithm (exact or heuristic) leading to good results is considered good, no matter if it can be mathematically justified. So the basic comparison is given below:

Table 1. Characteristics of tradition algorithm and WSRC

HAC,K-means,Bayesian	WSRC-needs
<ul style="list-style-type: none"> – Not necessarily fast – One-to-one association model – Stop-condition, or number of clusters to be found often needed 	<ul style="list-style-type: none"> – Not necessarily fast – One-to-one association model – Stop-condition, or number of clusters to be found often needed – Must be performed online – Overlapping clusters needed – Number or structure of topics not known in advance – Meaningful descriptions of clusters are required – Short, distorted input data (snippets) – Similarity criteria hard to define

Most clustering algorithms expect the data set to be clustered in the form of a set of vectors $1 \ 2 \ \{ \ , \ , \ } \ n \ X = \mathbf{x} \ \mathbf{x} \ \mathbf{K} \ \mathbf{x}$, where the vector, $1, \ , \ i \ \mathbf{x} \ i = \mathbf{K} \ n$ corresponds to a single object in the data set and is called the *feature vector*. Extracting the proper features to represent through the feature vector is highly dependent on the problem domain. If Feature Vector is quantitative then those algorithms are numerical and if this feature vector is in terms words then those called as phrase based. In text clustering algorithm two models are used heavily Suffix tree Document (STD) and Vector space model. The two main challenges with adapting clustering to suit the needs of web search engines and textual data has been to give good descriptive labels to the clusters and to be able to cluster documents on-the-fly in response to a particular user query. Traditional data mining approaches are not concerned with labeling clusters, but in return they are often very good at grouping data.

Unfortunately, regardless of how good the document grouping is, users are not likely to use a clustering engine if the labels are poor. Clustering performance is also a major issue, because web users expect fast response times. To deal with this linear time clustering algorithms that can cluster hundreds of documents in under a second have been developed. The requirements of web Search Results Clustering algorithm may be defined by:

- Identifying the structure of *topics* and creating *clusters* representing one or more topics, if their meaning can be considered close enough. If a hierarchical structure of *topics* can be identified, it should be mapped to the arrangement of clusters.
- The algorithm must work in a reasonably short time, so that the process of clustering is transparent to the user. Incremental and linear approaches are preferred over non-polynomial ones. P should be lower than N , because the objective of the algorithm is to identify documents sharing common topics, thus reducing the size of the results, not producing an alternative view over them.
- The algorithm must perform well with limited size of input data (or short snippets), or indicate that it is not possible to determine any logical structure of topics in the result set.
- The algorithm should be able to *describe* the topics forming a *cluster* in a manner intuitive for the user.

Table 2. Difference between numerical and phrase based algorithm

Numerical Algorithm	Phrase based algorithm
<ul style="list-style-type: none"> – Documents are converted to term document matrix – Numerical algorithms require more data than is available. – Raw numerical outcome is also difficult to convert back to cluster description – Data model is usually used Vector Space Model 	<ul style="list-style-type: none"> – Phrase-based on recurring phrases instead of numerical frequencies. – It is simple than numerical algorithms – These algorithms usually discards smaller clusters – Data model is usually used N-gram, Suffix Tree.

2.1.1 Stages of Web Search Results Clustering Algorithm

1) *Search Results Gathering*: The aim of this phase is to collect document snippets that have been returned by a search engine in response to the user's query. For web search results clustering the snippets can be obtained directly from a web search engine using an appropriate API, such as Google API, or by parsing the engine's HTML output and extracting all necessary data. A major issue in the gathering data of web search engine results is whether clustering the snippets returned by the search engines can produce results that are comparable to those achieved when clustering the corresponding web documents. Zamir et al [5] had studied and compared results of web documents and snippets and proved that snippets contain phrases that help in the correct clustering of the document, and do not contain some of the "noise" present in the original documents that might cause misclassification of the documents.

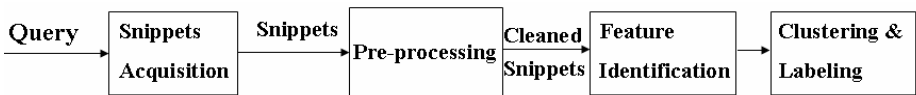


Fig. 2. Stages of clustering algorithm

2) *Preprocessing*: The process of searching a collection of documents⁸ involves a number of text processing techniques. Some of the techniques, such as the stemming, are essential parts of the process, whereas other, such as the lemmatization, wordnet processing are optional and are meant to increase the quality of search results. Natural language processing is another important domain in web search results clustering. There is lot of scope for improvement in nondeterministic formation of concept of web search results clustering. Eg. Almost all techniques use Porter Stemmer [12] which needs to be improved. Segond et al. [21] observed that part -of-speech tagging (POS) solves semantic ambiguity to some extent (40% in one of their tests). The author [18] stated that NER might serve a special purpose in combination of phrase based algorithm.

3) *Feature Identification*: The aim of the feature identification phase is to identify words in a text that are non-informative according to corpus statistics and can be omitted during clustering. The reasons for using feature selection/identification in search results clustering are twofold. First of all, in most cases limiting the number of features considerably increases the time efficiency of the clustering algorithm. Secondly, feature selection can help to remove noise from a text, which may result in higher accuracy of clustering. A large number of feature selection methods have been proposed in the literature, ranging from simple frequency thresholding to complex information theoretic algorithms. In numerical algorithms, term frequency, Document Frequency (DF), Term Strength (TS), Term Contribution (TC) are considered for feature identification. In some other algorithms words, phrases are considered. The different types of document representation models are vector space, n-gram model etc. In feature identification phase, first step is to represent the snippets to identify the words which are informative and can be used for clustering. Hence n-gram model will be given a preference.

4) *Clustering & Labeling*: The next step in the process chain is the actual clustering algorithm. Several classes of algorithms have been used for the search results clustering task, ranging from adaptations of the classic numerical approaches, such as K-Means in Scatter/Gather [13][20][22], to purpose-built methods such as Suffix Tree Clustering STC [1], Semantic On-line Hierarchical Clustering SHOC [6] and Tolerance Rough Set Clustering (TRC). We see the details of clustering approaches in next section. In the Label Discovery phase, each meaning group description is formed based on certain parameters.

2.2 Web Search Results Clustering Algorithm

There are various approaches for clustering like Neural Network, fuzzy logic, rough set, graph based etc. Web search clustering algorithms can be classified into two categories: numerical algorithms and phrase based algorithm. Strictly numerical algorithms require more data than available in a search results. Raw numerical outcome is also difficult to convert back to cluster description that human user understands. Phrase-based methods evolved to address this problem. We are studying phrase based and combination of both phrase based and numerical algorithms.

Cutting et al. [13] introduced document clustering as a document browsing method. They state that the Scatter/Gather system [14] [15] is particularly helpful in situations in which it is difficult or undesirable to specify a query formally i.e. when the user is not looking for anything specific, just wants to discover the general information content of the corpus (to gain an overview); or when it is difficult to formulate the query precisely (help user formulate a search request). Two near linear time clustering algorithms were presented: Buckshot algorithm and Fractionation algorithm. However, their work is based on general document collections, not on dynamically generated search results. Zamir & Etzioni [5] followed this paradigm and proposed the notion of search results clustering. In their Grouper system, STC (Suffix Tree Clustering) treats a document as a string instead of a set of words. It attempted to cluster documents “snippets” returned by search engine according to common phrases they contain, thus employing information about the proximity and order of single keywords in addition to their frequencies. STC has two key features: the use of phrases and a simple cluster definition that does not assume a specific model for the data. This algorithm made use of special data structure (which is a kind of inverted index of phrases for document collection) that can be constructed incrementally and linearly in time.

Several characteristics make STC a promising candidate for the clustering of search results. First, it is phrase-based, generating clusters by grouping documents that share many phrases. Phrases are also useful in constructing concise and accurate descriptions of the clusters. Second, it does not adhere to any model of the data. Its only assumption is that documents on the same topic will share common phrases. Third, STC allows overlapping clusters. It is important to avoid confining each document to only one cluster since documents often have multiple topics, and thus might be similar to more than one group of documents. Fourth, STC uses a simple cluster definition – all documents containing one of the cluster’s phrases are members of the cluster. Finally, STC is a fast incremental, linear time (in the number of documents) algorithm, which makes it suitable for online clustering of Web searches.

The Suffix Tree Clustering algorithm works in two main phases: base cluster discovery phase and base cluster merging phase. In the first phase, a generalized suffix tree of all texts’ sentences is built using words as basic elements. After all sentences are processed, the tree nodes contain information about the documents in which particular phrases appear. Using that information, documents that share the same phrase are grouped into base clusters of which only those are retained whose score exceeds a predefined Minimal Base Cluster Score. In the second phase of the algorithm, a graph representing relationships between the discovered base clusters is built based on their similarity and on the value of the Merge Threshold. Base clusters belonging to coherent sub graphs of that graph are merged into final clusters

A clear advantage of Suffix Tree Clustering is that it uses phrases to provide concise and meaningful descriptions of groups. However, as noted in STC, thresholds play a significant role in the process of cluster formation, and they turn out particularly difficult to tune. Also, STC’s phrase pruning heuristic tends to remove longer high-quality phrases, leaving only the less informative and shorter ones. Finally, as pointed out in, if a document does not include any of the extracted phrases it will not

be included in the results although it may still be relevant. As STC algorithm is baseline for online web search result clustering algorithms, threshold there are many attempt to improve STC algorithm. Suffix tree with new scoring function [28] is basically Extension of STC, specifically new scoring method, is applied and hierarchical clustering method is used so that we can tune clusters which gets merged.

Extended suffix tree clustering (ESTC)[4] introduces a new cluster scoring function and a new cluster selection algorithm to overcome the problems with overlapping clusters. It significantly improves the results.

Table 3. Difference between numerical and phrase based algorithm

Suffix tree Clustering	Suffix Tree with new scoring function	Extended Suffix tree Clustering
Step 1: Cleaning Stemming, Sentence boundary identification. Punctuation Elimination.	Step 1: Cleaning Stemming, Sentence boundary identification. Punctuation Elimination.	Step 1: Cleaning Stemming, Sentence boundary identification. Punctuation elimination.
Step 2: Suffix tree construction for each node in the tree { if (number of documents in node's subtree > 2) { if (candidate Base Cluster Score > Minimal Base Cluster Score) { add a base cluster to the list of base clusters;}}}	Step 2: Suffix tree construction for each node in the tree { if (number of documents in node's subtree > 2) { if (candidate Base Cluster Score > Minimal Base Cluster Score) { add a base cluster to the list of base clusters;}}}	Step 2: Suffix tree construction Build suffix tree. Produces base clusters (internal nodes) Base cluster score= $ D $ documents. Merge Cluster score of document is average of base clusters
Step 3 : Merging Build a graph where nodes are base clusters and there is a link between node A and B if and only if the number of common documents indexed by A and B is greater than the Merge-threshold; clusters are coherent subgraphs of that graph;	Step 3 : Merging Hierarchical merging based on score based on single link property. Priority queue is used from which highest score two documents get pops and get merged.	Step3 :Cluster Selection Algorithm for each k-step { selecting the best cluster (with some heuristic) at each step corresponds to a greedy search, which is equivalent to using 0-step look-ahead } Branch and bound Pruning is used for further merging.

Carrot system built by Weiss and Stefanowski [10] [11] [16], extended STC's application into the Polish Language, by using different stemming techniques. They investigated the influence of two primary STC parameters: merge threshold and minimum base cluster score, on the number and quality of results produced by STC algorithm.

Semantic Hierarchical On-line Clustering SHOC [6] is based on latent semantic indexing and designed to work in Chinese. Two novel concepts are introduced to overcome STC's limitations: complete phrases and continuous cluster definition. A data structure called suffix array is used to identify complete phrases to avoid extracting meaningless partial phrases. Continuous cluster definition allows documents to be assigned to multiple clusters.

LINGO [11] is a slightly modified version of the SHOC algorithm and is claimed as a "description oriented algorithm". Being different from the previous approach, LINGO identifies cluster labels first using latent semantic indexing technique, retrieved search results are assigned to different groups based on the labels.

Table 4. Difference between SHOC and LINGO

	SHOC	LINGO
Preprocessing	Does not include language identification step	Includes snippet language identification step
Phrase identification	Based on Suffix Arrays	Adapted from SHOC
Label discovery	Performed <i>after</i> cluster content Discovery	SVD, performed <i>before</i> cluster content discovery
Cluster content discovery	Based on Singular Value Decomposition	Based on cluster labels, employs the Vector Space Model
Post-processing	Hierarchical cluster merging	No cluster merging applied
Advantages	It uses Complete Phrase and Continuous Phrase Suffix array is used instead STC.	Readable cluster label Diverse cluster label Overlapping clusters Ease of tune
Disadvantages	Thresholds are used and it is difficult to select their value SVD produce unintuitive continuous cluster	Flat Cluster, Fixed number of clusters, "Other Topic" problem Over specified cluster Label Computationally expensive (SVD computationally high) Lack of incremental processing

Jiang et al [23] developed the Retriever system, comparing two different distance metrics: N-gram and Vector space model by clustering the data using a robust fuzzy relational algorithm. They compare the results with STC and find that the N-gram based approach performs better than the vector space based approach [17], and also Zamir and Etzioni's STC algorithm. But the author pointed out that their search results are drawn from Lycos, whereas STC draws on Metacrawler. So the comparison between these two methods is arguable.

STC generates too many clusters which is not very useful to the user as user cannot possibly handle so many clusters. To address overlap documents Extended Suffix Tree Clustering (ESTC) was introduced by [4]. They have extended their work and introduced new algorithm called as Query directed web page clustering (QDC) [19] which is not based on suffix tree but based on term frequency.

The original STC algorithm can often construct a long path of suffix tree, particularly when the same snippets are fed to the STC algorithm. Hau-Jun Zeng and etc. [8] introduced an improved suffix tree with n-gram to deal with the problem of the original STC algorithm. However, the suffix tree with n-gram can discover only partial common phrases when the length of n-gram is shorter than the length of true common phases.

Jongkol Janruang et al [3] have proposed new approach for web search result clustering to deal with such problems. The new approach still uses a suffix tree with n-gram. However, the approach also introduces a new base cluster combining technique with a new partial phase join operation to find a true common phase.

There are some algorithms proposed in year recent which used latent semantic indexing and Singular Value Decomposition (SVD). Giansalvatore Mecca et al [7] proposed new algorithm based on dynamic SVD. The features of this algorithm are that it uses Latent Semantic Indexing technique on whole document with a basic aim to increase the performance of SVD to discover the optimal number of singular values that can be used for clustering purpose.

As we have seen in Section A and Table 4 there are two data models used in clustering VSD (Vector Space model) and STD (Suffix Tree Document). We have just summarized the algorithms and their respective data models used.

Table 5. Difference between STD Model and VSD based algorithm

STD model based Algorithms	VSD based Algorithms
– Suffix tree Clustering	– SHOC
– Extended suffix tree Clustering	– LINGO
– A true common phrase label Discovery algorithm	– Dynamic SVD

The STD model, we find that this model can provide a flexible n-gram method to identify and extract all overlap phrases. Despite of all work on suffix tree based algorithm [1] [6] [4], an effective method has not been found to evaluate the effect of each phrase in clustering algorithm.

In contrast the numerical algorithm based on VSD models uses feature vector to represent a document. The statistical features of all words are taken into account of the term weights (usually tf-idf) and similarity measures, whereas the sequence order of words is rarely considered in the clustering approaches based on the VSD model. Both VSD and STD play more important role text based clustering. Here we have compared some recent algorithms based on theoretical analysis which are given in Appendix.

3 Summary and Future Direction

Being a fairly young discipline of computer science, web search results clustering is gaining increasingly promising perspectives for the future. On the one hand, extensive research is being conducted on faster and more precise clustering algorithms, and on the other, the scientific advances are followed by practical implementations of the web search clustering idea, some of which, such as Vivisimo, have already achieved commercial success. With this paper we aim to contribute to the scientific trend in web search clustering. We present an overview of web search results clustering approaches. As the search results are retrieved from a (meta) search engine dynamically, it introduces lots of challenges and makes ephemeral clustering very different from traditional document clustering. There are some problems that have not been well addressed yet. There are many research directions to follow and many bold ideas waiting to be implemented. The obvious goal should be to improve the quality of clusters, their *informative* value to the user of a search engine. A very important issue here is to find some objective measure of user satisfaction; otherwise it will be hard to assess the field's progress. Even if new algorithms are hard to come up with, existing methods have a number of drawbacks one can work on. For example, creating a better cluster merging method in STC is necessary, also further experiments on using bags of words instead of ordered phrases seem very promising. In the light of search engines' popularity regardless of their mentioned drawbacks, WSRC has a key role to play: bringing new quality tools and facilitating searching for information in the internet. This mission can be accomplished only with proper understanding of the problem and new, specialized algorithms. Here with we give the future direction:-

1). We have come to the conclusion that combination of Phrase based and Numerical Algorithm will give better results. 2). There are some author quotes that, there is

enormous influence of preprocessing phase on the overall quality. 3). There is also possible improvement in presenting search results and for that hierarchical clusters are required. 4). Phrase identification is a very important phase in Phrase based algorithm. There is a need for improvement in phrase extraction and the identification of a proper similarity measures. 5) Experiments on a large scale are required for subjective evaluation of the algorithm. 6). There are very few algorithms available currently that provide proper labels. Improvement in labeling techniques will facilitate better representation of the search engine on short screen like mobiles.

References

1. Zamir, O., Etzioni, O.: Web Document Clustering: A Feasibility Demonstration. In: Proc. ACM SIGIR 1998, Melbourne, Australia, pp. 46–54 (1998)
2. Zhong, S.: Semi-supervised model-based document clustering: A comparative study. Springer, Heidelberg (2006)
3. Janruang, J., Kreesuradej, W.: A New Web Search Result Clustering based True Common Phrase Label Discovery. In: International Conference on Computational Intelligence for Modeling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC 2006) (2006)
4. Crabtree, D., Gao, X., Andreae, P.: Improving Web Clustering by Cluster Selection. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2005 (2005)
5. Zamir, O.: Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results. PhD Thesis University of Washington (1999)
6. Zhang, D., Dong, Y.: Semantic, Hierarchical, Online Clustering of Web Search Results (unpublished)
7. Mecca, G., Raunich, S., Pappalardo, A.: A new algorithm for clustering search results. *Science Direct Data & Knowledge Engineering* 62, 504–522 (2007)
8. Zeng, H.-J., et al.: Learning to Cluster Web Search Results. In: SIGIR 2004, Peking University (2004)
9. Campos, R., Dias, G., Nunes, C.: WISE: Hierarchical Soft Clustering of Web Page Search Results based on Web Content Mining Techniques. In: Proceedings of the 2006. IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006). IEEE, Los Alamitos (2006)
10. Osinski, S., Weiss, D.: Conceptual clustering using lingo algorithm: Evaluation on open directory project data. In: IIPWM 2004 (2004)
11. Osiński, S.: An algorithm for clustering of web search results, Master thesis (2003)
12. Porter, M.F.: An algorithm for suffix stripping. In: *Readings in Information Retrieval*, pp. 313–316 (1997)
13. Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W.: Scatter/gather: a cluster-based approach to browsing large document collections. In: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 318–329 (1992)
14. Pirolli, P., Schank, P., Hearst, M., Diehl, C.: Scatter/gather browsing communicates the topic structure of a very large text collection. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 213–220 (1996)

15. Hearst, M.A., Pedersen, J.O.: Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In: Proceedings of SIGIR 1996, 19th ACM International Conference on Research and Development in Information Retrieval, Zürich, CH, pp. 76–84 (1996)
16. Weiss, D.: Carrot2 Developers Guide, <http://www.cs.put.poznan.pl/dweiss/carrot/site/developers/man>
17. Wroblewski, M.: A hierarchical www pages clustering algorithm based on the vector space model. Master's thesis, Poznan University of Technology, Poland (July 2003)
18. Borch, H.O.: Clustering On-line Clustering of Web Search Results. M.S. Thesis Norwegian University of Science and Technology (2006)
19. Crabtree, D., Gao, X., Andreae, P.: Improving Query Directed Web Page Clustering. In: Proceedings of the 2006 (2006)
20. Ferragina, P., Gulli, A.: The Anatomy of a Hierarchical Clustering Engine for Webpage, News and Book Snippets. Technical report, RR04-04 Informatica, Pisa (2004)
21. Segond, F., Shiller, A., Grefenstette, G., Chanod, J.-P.: An Experiment in Semantic Tagging Using Hidden Markov Model Tagging. In: Dans ACL 1997 Workshop on Information Extraction and the Building of Lexical Semantic Resources for NLP Applications (1997)
22. Ferragina, P., Gulli, A.: A personalized Search Engine Based On Web-Snippet hierarchical Clustering. In: 14th International World Wide Web Conference (2005)
23. Jiang, Z.H., Joshi, A., Krishnapuram, R., Yi, L.Y.: Retriever: improving web search engine results using clustering. In: Managing Business and Electronic Commerce (2002)
24. Wang, Y., Zuo, W., Peng, T., He, F., Hu, H.: Clustering Web Search Results Based on Interactive Suffix Tree Algorithm. In: Third 2008 International Conference on Convergence and Hybrid Information Technology (2008)
25. Wen, H., Huang, G.-S., Li, Z.: Clustering web Search using semantic information. In: Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, July 12-15 (2009)
26. Chim, H., Deng, X.: Efficient Phrase-based Document similarity for clustering. IEEE Transaction on Knowledge and Data Engineering 20(9) (September 2008)
27. Kale, A., Bharambe, U.: A New Suffix Tree Similarity Measure and Labeling for Web Search Results Clustering. In: Second International Conference on Emerging Trends in Engineering & Technology, pp. 856–861 (2009)
28. <http://www.stanford.edu/class/cs276a/projects/.../arigreen-sbranson.pdf> (visited_on march 2009)

Appendix

Steps	Dynamic SVD[7]	A true common phrase discovery algorithm[3]	Hierarchical soft clustering technique[9]	QDC[19]
Search Results Gathering	Not mentioned in algorithm. (Assumes that results are gathered from top rank search result)	Assumes that results are gathered from search engines	'Select algorithm' which selects most relevant documents over the set of web pages	Assumes that results are gathered from search engines
Preprocessing:	Document representation using Vector Space Model and stemming and stop word removal is used	By stemming Non-word tokens removal is used	By a new method of phrase extraction from relevant web pages	Standard preprocessing technique.
Feature Identification	Document representation using Vector Space Model and Based on incremental SVD and LSI	Building suffix tree with n-gram	Extracting the phrases using SENTA based on 3 concepts (Positional n-gram, the association measure Mutual Expectation and the GenLocalMaxs algorithms).	Words and its frequencies.
Label Discovery	Labels are the frequently occurring terms.	Concatenation of the edge-labels on the path from the root to that node.	Using a heuristic that chooses commonly occurring key concept.	Nothing specific is given in paper
Advantages	It considers statistical properties of words, uses VSD model The first algorithm which introduced Dynamic SVD and it is incremental.	It is completely phrase based algorithm It uses both n-gram and suffix tree clustering approaches. It generates label automatically. It improves the clusters by ranking.	Uses WISE(a meta search engine), that automatically builds clusters of related web pages embodying one meaning of the query and also provide a topic and language independent real-world web. It gives quality results through an organized and disambiguated structure of concepts.	Introduces a new similarity measure NGD(normalized google distance). It identifies better clusters using a query directed cluster quality guide it fixes the cluster chaining (drifting) problem using a new cluster splitting method. It uses heuristic for cluster quality and it improves the clusters by ranking the pages according to cluster relevance.
Disadvantages	Complex computations	Clusters Overlap, quality relevance is not considered	It is based Web page clustering where we required a whole document for clustering which will decrease the performance.	It is just single word based algorithm. Multiword unit phrases are not considered.

Steps	Clustering using semantic information [26]	ISTC[24]	Efficient phrase based document similarity for clustering[26]	A New Suffix Tree Similarity Measure and Labeling for WSRC[27]
Search Results gathering	Snippets from search engine	Snippets from google	Web documents	Snippets from yahoo search engine.
Preprocessing	Porter Stemmer, Stop word punctuation, Non token removal	Normal preprocessing technique is used.	Assumed to normal to be normal preprocessing.	Punctuation, stopword and porter stemmer is used.
Feature Identification	Generalized suffix tree is used.	To reduce time complexity it introduced Link list based improved suffix tree data structure.	Generalized suffix tree	Suffix tree with n-gram
Clustering	Clustering using new scoring function, similarity measure which is based on union set and for selecting final clusters SVD and Latent semantic indexing is used	Traversing the tree based on certain condition clustering is performed.	Suffix tree is mapped with VSD and then cosine similarity is used.	For merging base clusters new algorithm is proposed which is based on cosine similarity and phrase overlap.
Labeling	A new mathematical formula is used based on length of phrase and statistical properties of term.	Cluster Label is extracted during suffix tree traversal and threshold (less than number of documents in which the phrase appears) is defined.	Not mentioned	Labeling on frequency of particular term in that cluster.
Advantages	It is combination of STD and VSD base algorithm. It is semantic based labeling.	Cluster label is more readable, concise and applicable to the category. Overlapping for cluster Interactive clustering.	It introduced group –average hierarchical clustering algorithm with phrase-based document similarity	Successful combination of text data clustering and numerical data clustering.
Disadvantages	It is based on threshold. It is evaluated on less users.	Recursive algorithm. Semantic feature is not considered.	Semantic of word is not considered	Threshold plays important role Semantic of word is not considered.

An Improved K-Means Clustering Approach for Teaching Evaluation

Oswal Sangita and Jagli Dhanamma

Department of MCA
Vivekanand Education Society's Institute Of Technology
University of Mumbai, Mumbai-71, India
sangita_oswal1@rediffmail.com
dhana1210@yahoo.com

Abstract. Intelligent evaluation as an important branch in the field of artificial intelligence is a decision-making process of simulating the domain experts to solve complex problems. In this paper, we put forward a kind of intelligent evaluation method based on clustering, which can be used to mine different groups of teachers and evaluate the teaching quality automatically. Clustering analysis method is one of the main analytical methods in data mining, which influences the clustering results directly. In this paper Firstly, we do some improvement on traditional K-means clustering due to its shortcomings. Secondly, we propose a model or teaching quality evaluation based on improved K-means clustering.

Keywords: Teaching Evaluation; Clustering; K-Means; Data mining.

1 Introduction

Clustering is a way that classifies the raw data reasonably and searches the hidden patterns that may exist in datasets [7]. It is processes of grouping data objects into disjointed clusters so that the data's in the same cluster are similar, yet data's belonging to different cluster differ. The aim of cluster analysis is exploratory, to find if data naturally falls into meaningful groups with small within-group variations and large between-group variation. Clustering methods only try to find an approximate or local optimum solution. The demand for organizing the sharp increasing data's and learning valuable information from data, which makes clustering techniques are widely applied in many application areas such as artificial intelligence, biology, customer relationship management, data compression, data mining, information retrieval, image processing, machine learning, marketing, medicine, pattern recognition, psychology, statistics[3].

The data mining in teaching evaluation is that applying the appropriate ways to extract, convert, analyze and process the data from the teaching-related database, in order to make critical decision [5]. The major mining methods of teaching quality evaluation are dynamic prediction, association analysis and clustering [6]. The main idea of the clustering technique is to group all nodes into multiple separated entities called clusters, and consequently form a cluster structure. With the purpose of doing teaching evaluation intelligently and accurately, we introduce clustering into the process of teaching quality Evaluation. In this paper, we propose an intelligent evaluation

model based on clustering algorithm, which can be easily applied to do evaluation in the domain of adult higher education. Clustering is a challenged research field which belongs to unsupervised learning. The number of clusters we needed is unknown and the formation of clusters is data driven completely. Clustering can be the pretreatment part of other algorithms or an independent tool to obtain data distribution, and also can discover isolated points.

K-means is a numerical, unsupervised, non-deterministic, iterative method. It is simple and very fast, so in many practical applications, the method is proved to be a very effective way that can produce good clustering results. But it is very suitable for producing globular clusters. Several attempts were made by researchers to improve efficiency of the k-means algorithms [5]. In the literature [3], there is an improved k-means algorithm based on weights. This is a new partitioning clustering algorithm, which can handle the data of numerical attribute, and it also can handle the data of symbol attribute. Meanwhile, this method reduces the impact of isolated points and the “noise”, so it enhances the efficiency of clustering. However, this method has no improvement on the complexity of time. In the literature [1], it proposed a systematic method to find the initial cluster centers, this centers obtained by this method are consistent with the distribution of data. Hence this method can produce more accurate clustering results than the standard k-means algorithm, but this method does not have any improvements on the executive time and the time complexity of algorithm.

This paper presents an improved k-means algorithm. Although this algorithm can generate the same clustering results as that of the standard k-means algorithm, the algorithm of this paper proposed is superior to the standard k-means method on running time and accuracy, thus enhancing the speed of clustering and improving the time complexity of algorithm. By comparing the experimental results of the standard k-means and the improved k-means, it shows that the improved method can effectively shorten the running time.

2 K-Means Clustering

K-means is the simplest and most popular classical clustering method that is easy to implement. The classical method can only be used if data about all objects is located in main memory. The method is called K-means since each of the K clusters is represented by the mean of the objects called the centroid within it. It is also called centroid method since at each step the centroid point of each cluster is assumed to be known and each of the remaining points are allocated to the cluster whose centroid is closest to it. Once this allocation is completed the centroid of the cluster are recomputed using simple means and process of allocating points to each cluster is repeated until there is no change in the clusters .

a) The process of k-means algorithm

This part briefly describes the standard k-means algorithm. K-means is a typical clustering algorithm in data mining and which is widely used for clustering large set of data. In 1967, MacQueen firstly proposed the k-means; it was one of the most simple, non-supervised learning algorithms, which was applied to solve the problem of the well-known cluster [2].

The K-means method uses the Euclidean distance measure, which appears to work well with compact clusters. The Euclidean distance between one vector $x=(x_1, x_2, \dots, x_n)$ and another vector $y=(y_1, y_2, \dots, y_n)$, can be obtained as follow:

$$D(x_i, y_i) = [\sum (x_i - y_i)^2]^{1/2}$$

The process of k-means algorithm as follow:

1. Select the number of clusters let this number be K .
2. Pick K seed as centroids of the K clusters. The seed may be picked randomly unless the user has some insight into the data.
3. compute the Euclidean distance of each object in the data set from each of the centroids
4. Allocate each object to the cluster it is nearest to based on the distances computed in the previous step.
5. Compute the centroids of the clusters by computing the means of the attribute values of the objects in each cluster.
6. Check if stopping criteria has been met, if not go to step 3.

The k-means clustering algorithm always converges to local minimum. Before the k-means algorithm converges, calculations of distance and cluster centers are done while loops are executed a number of times, where the positive integer t is known as the number of k-means iterations. The precise value of t varies depending on the initial starting cluster centers [8]. The distribution of data points has a relationship with the new clustering center, so the computational time complexity of the k-means algorithm is $O(nkt)$, where n is the number of all data objects, k is the number of clusters; t is the iterations of algorithm. Usually requiring $k \ll n$ and $t \ll n$.

b) The shortcomings of k-means algorithm

We can see from the above analysis of algorithms, the algorithm has to calculate the distance from each data object to every cluster center in each iteration. However, by experiments we find that it is not necessary for us to calculate that distance each time. Assuming that cluster C formed after the first j iterations, the data object x is assigned to cluster C , but in a few iterations, the data object x is still assigned to the cluster C . In this process, after several iterations, we calculate the distance from data object x to each cluster center and find that the distance to the cluster C is the smallest. So in the course of several iterations, k-means algorithm is to calculate the distance between data object x to the other cluster center, which takes up a long execution time thus affecting the efficiency of clustering.

3 Improved K-Means Clustering Algorithm

The standard k-means algorithm needs to calculate the distance from the each date object to all the centers of K Clusters when it executes the iteration each time, which takes up a lot of execution time especially for large-capacity databases. For the shortcomings of the above k -means algorithm, this paper presents an improved k-means method. The main idea of algorithm is to set two simple data structures to retain the labels of cluster and the distance of all the date objects to the nearest cluster during

the each iteration, that can be used in next iteration, we calculate the distance between the current data object and the new cluster center, if the computed distance is smaller than or equal to the distance to the old center, the data object stays in it's cluster that was assigned to in previous iteration. Therefore, there is no need to calculate the distance from this data object to the other $k-1$ clustering centers, saving the calculative time to the $k-1$ cluster centers. Otherwise, we must calculate the distance from the current data object to all k cluster centers, and find the nearest cluster center and assign this point to the nearest cluster center. And then we separately record the label of nearest cluster center and the distance to its center. Because in each iteration some data points still remain in the original cluster, it means that some parts of the data points will not be calculated, saving a total time of calculating the distance, thereby enhancing the efficiency of the algorithm.

The process of the improved algorithm is described as Follows:

- 1) Randomly select K objects from dataset D as initial cluster centers.
- 2) Calculate the distance between each data object d_i ($1 \leq i \leq n$) and all K cluster centers c_j ($1 \leq j \leq k$) as Euclidean distance $d(d_i, c_j)$ and assign data object d_i to the nearest cluster.
- 3) For each data object d_i , find the closest center c_j and assign d_i to cluster center j ;
- 4) Store the label of cluster center in which data object d_i is and the distance of data object d_i to the nearest cluster and store them in array Cluster[] and the Dist[] separately.

Set Cluster[i] =j, j is the label of nearest cluster. Set Dist[i] =d (d_i, c_j), d (d_i, c_j) is the nearest Euclidean distance to the closest center.

- 5) For each cluster j ($1 \leq j \leq k$), recalculate the cluster center;
- 6) Repeat
- 7) For each data object d_i Compute it's distance to the center of the present nearest cluster;
 - a. If this distance is less than or equal to Dist[i], the data object stays in the initial cluster;
 - b. Else For every cluster center c_j ($1 \leq j \leq k$), compute the distance $d(d_i, c_j)$ of each data object to all the center, assign the data object d_i to the nearest center c_j . Set Cluster[i]=j; Set Dist[i]= $d(d_i, c_j)$;
- 8) For each cluster center j ($1 \leq j \leq k$), recalculate the centers;
- 9) Until the convergence criteria is met.
- 10) Output the clustering results;

The improved algorithm requires two data structure (Cluster [] and Dist []) to keep the some information in each iteration which is used in the next iteration. Array cluster [] is used for keep the label of the closest center while data structure Dist [] stores the Euclidean distance of data object to the closest center. The information in data structure allows this function to reduce the number of distance calculation required to assign each data object to the nearest cluster, and this method makes the improved k-means algorithm faster than the standard k-means algorithm.

Firstly, this paper proposes an improved k-means algorithm, to obtain the initial cluster, time complexity of the improved k-means algorithm is $O(nk)$. Here some data points remain in the original clusters, while the others move to other clusters. If the

data point retains in the original cluster, this needs $O(1)$, else $O(k)$. With the convergence of clustering algorithm, the number of data points moved from their cluster will reduce. If half of the data points move from their cluster, the time complexity is $O(nk/2)$. Hence the total time complexity is $O(nk)$. While the standard k-means clustering algorithm require $O(nkt)$. So the proposed k-means algorithm in this paper can effectively improve the speed of clustering and reduce the computational complexity. But the improved k-means algorithm requires the predestinated number of clusters, k , which is the same to the standard k-means algorithm. If you want to get to the optimal solution, you must test the different value of k . Secondly, this paper is describing model for teaching evaluation based on improved k-mean algorithm.

4 The Model of Intelligent Evaluation

Teaching evaluation is based on the certain educational principles and policies. We propose the targeted advice process of the further improvement and development.

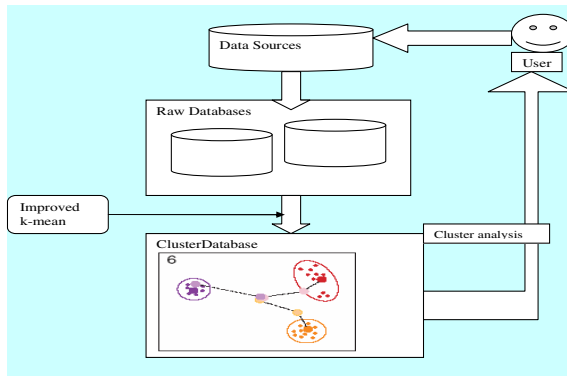


Fig. 1. The Teaching Evaluation Model based on an Improved K-Means Clustering

Teaching evaluation system is the platform that the school leaders and managers make use of to master the teaching goals and improve the quality of the teaching. Based on the school network platform, we make the design method of the teaching evaluation system. For the characteristics of the adult higher education school database, this paper presents the education quality assessment framework based on an improved k mean cluster algorithm. The model consists mainly of data collection, data selection, data standardization and rules recommended, as shown in Fig. 1.

a) Data Acquisition

The goal of the data acquisition is to achieve the integrity of the large-scale evaluation data, and accurately to collect the data. Teaching quality analysis and evaluation system uses the computer's data conversion function to directly open source database for data collection. We integrate the various types of data from transaction-oriented real-time operation database to data warehouse based on the data mining analysis, mainly

including the students' evaluation information, teacher information, evaluation criteria and class teachers' instructor information integration.

b) Cluster formation

We extract the property from the teachers' archive database. And we merge it with the students' assessment results to group a table. That provides data mining objects for the entire data mining module. As the improved k-mean cluster algorithm is fit for the data mining of the untrained database. We make use of the data mining technology to analyze the collected sample data. We try to discover useful knowledge (cluster) hidden in the data and extract the useful knowledge for the schools and teachers.

c) Cluster Analysis

After applying the improved K-mean on the sample data set we get the pattern for group of teachers.

Green color represents cluster C1 (Above Average)

Red color represents cluster C2 (Average)

Blue color represents cluster C3 (Below Average)

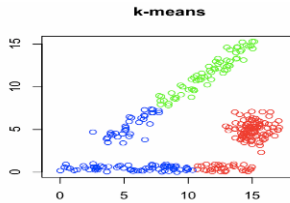


Fig. 2. Cluster generated

5 Case Study

In order to prove the correctness and effectiveness of the model proposed in this paper, this section will validate the method through a real-life example. The goal of this test is to mine and analysis evaluated data information based on the existing quality of education in order to provide basis of decision-making for improving the teaching quality of teachers and the school's overall standard of teaching. As affected by different factors, different survey objects share different weight in teaching quality evaluation grades, so that the different survey objects have various weighting in teachers teaching quality assessment.

Table 1.

Fid	Questioner	
	Student	Expert
1	45	45
2	35	35
3	49	49
.....
10	35	39

Table 2.

Fid	Inspection Standard				
	Teaching attitude	Teaching method	Teaching level	Teaching effect	Language
1	15	17	18	19	19
2	12	12	12	12	12
3	19	16	15	17	18
-	---	----	-----	-----	-----
10	15	13	12	15	15

Table 3.

Fid	Age	Education Qualification	Assessment Weight	Faculty score	Allocation To nearest Cluster
1	45	5	90	88	C1
2	35	4	70	60	C3
3	33	4	98	85	C1
...	
10	25	3	74	70	C2

Table 3 shows the teachers' score calculated by the parameters from Table 1 and Table 2. Assessment weight related to Questioner Faculty score related to Inspection standard.

6 Conclusion

K-means is a typical clustering algorithm and it is widely used for clustering large sets of data. This paper elaborates *k*-means algorithm and analyses the shortcomings of the standard *k*-means clustering algorithm. Because the computational complexity of the standard *k*-means algorithm is objectionably high owing to the need to reassign the data points a number of times during every iteration, which makes the efficiency of standard *k*-means clustering is not high. This paper presents a simple and efficient way for assigning data points to clusters. The proposed method in this paper ensures the entire process of clustering in $O(nk)$ time without sacrificing the accuracy of clusters. In this paper, we suggested a model of teaching quality Evaluation based on a refined *k*-means clustering including its analyzing and design. We described the data acquirement, cluster formation and analysis on how to obtain a useful teaching evaluation data and created an appropriate cluster.

References

1. Yuan, F., Meng, Z.H., Zhang, H.X., Dong, C.R.: A New Algorithm to Get the Initial Centroids. In: Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26–29 (August 2004)

2. Sun, J., Liu, J., Zhao, L.: Clustering algorithms Research. *Journal of Software* 19(1), 48–61 (2008)
3. Sun, S., Qin, K.: Research on Modified k-means Data Cluster Algorithm. In: Jacobs, I.S., Bean, C.P. (eds.) *Fine Particles, Thin Films and Exchange Anisotropy*, *Computer Engineering*, vol. 33(13), pp. 200–201 (July 2007)
4. Xie, Q.L.: Data Mining in Teaching Quality Evaluation Based on Association Rules. *Modern Computer* (6) (2008)
5. Yuan, Y., Li, H.: Research on Teaching Evaluation System Based on Data Mining. *Computer and Modernization* (11) (2009)
6. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery* 2, 283–304 (1998)
7. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases
8. Abdul Nazeer, K.A., Sebastian, M.P.: Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. In: *Proceeding of the World Congress on Engineering*, London, vol. 1 (July 2009)

Consensus Based Dynamic Load Balancing for a Network of Heterogeneous Workstations

Janhavi Baikerikar¹, Sunil Surve², and Sapna Prabhu²

¹ Postgraduate Student

² Assistant Professor

Fr.Conceicao Rodrigues College of Engineering, Bandra West, Mumbai, India

janhavibaikerikar@gmail.com,

surve@frcrce.ac.in, sapna@frcrce.ac.in

Abstract. In this paper, we propose consensus based load balancing algorithm for network of heterogeneous workstations. Load balancing is a critical issue in parallel and distributed computing in order to carry out efficient utilization of computational resources. The focus of this work is to design a dynamic decentralized load balancing algorithm for a network of heterogeneous workstations using consensus theory and graph partitioning. Simulation results show that the proposed load balancing algorithm is scalable, reliable, fault tolerant and maintains a balance between communication overhead and load balancing time.

Keywords: Consensus, Graph partitioning, Load balancing.

1 Introduction

In a distributed environment such as grid or cluster the workstations are connected in an arbitrary topology. Now even though we have many workstations connected together, we come across many difficulties that we did not encounter earlier.

One of these difficulties is the distribution of workload among these workstations. Workload is set of jobs or programs which are independent of one another. Work load distribution is the task of distributing a set of jobs among all workstations of a given network. This distribution leads to some workstations getting totally utilized while other remaining unused. This creates a situation of uneven load distribution. Such a solution is not feasible in the distributed environment. This situation is avoided by distributing the load evenly among all the workstations such that none of the workstations is over loaded or under loaded [1]. This is called load balancing (figures 1 and 2).

In this paper, we present the consensus based dynamic load balancing algorithm for a network of heterogeneous workstations. The algorithm developed is scalable, reliable and fault tolerant. It maintains a balance between the communication overhead and load balancing time.

This paper is organized as follows. In section 2 we discuss some of the current algorithms for load balancing. This is followed by basics of graph theory and splitting of graph, the information exchange and definition of the consensus parameter acceptance value in section 3. Simulation results are given in section 4. We conclude in the last section.

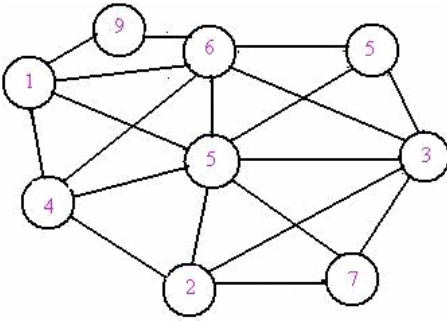


Fig. 1. Jobs on each workstation before load balancing

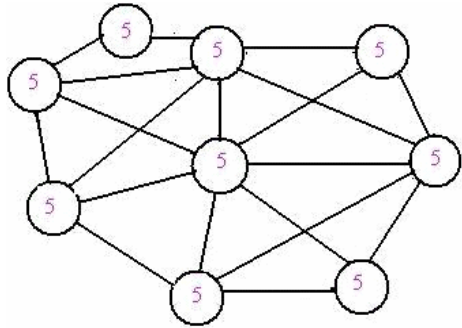


Fig. 2. Jobs on each workstation after load balancing

2 Background

Conceptually, load balancing algorithms can be classified into two categories: *static* or *dynamic* [2]. Static load balancing algorithms assume that a priori information about all the characteristics of the jobs, the workstations and the communication network are known and provided. Load balancing decisions are made at compile time and remain constant during runtime. The static load balancing algorithms do not take the load on individual workstation into consideration. The static approach is attractive because of its simplicity and the minimized runtime overhead. The dynamic load balancing algorithms perform at runtime but at the cost of increased communication overhead. The aim of a good dynamic decentralized load balancing approach is to maintain a balance between load balancing time and communication overhead. In this section we give brief overview of the current dynamic load balancing algorithms.

Abed et al, [3] presented the Bidding Approach in which the bid value is the criteria for scheduling the jobs inside the network. The major limitation of this approach is that at a given time only 2 workstations can participate in load balancing. Shah et al, [4] proposed MELISA (Modified Estimated Load Information Scheduling Algorithm) algorithm in which the job migration cost is primary factor for distributing jobs inside a predefined partition.

Venu Gopalachari et al, [5] presented centralized Dynamic Scheduling algorithm which considers the load of each cluster and the load of each workstation in the cluster, selects a cluster with appropriate weight and allocates the jobs in a cluster according to the load of the workstations. P. Chandra et al, [6] have proposed a centralized load-balancing scheme based upon system heterogeneity in which I/O-intensive jobs are migrated to the fastest workstation. This load balancing scheme minimizes the average slow down of all parallel jobs running on a cluster and reduces the average response time of the jobs.

3 Consensus Based Dynamic Load Balancing for a Network of Heterogeneous Work Stations

We represent the network as a graph. This is then followed by partitioning the graph into partitions. Each workstation calculates its acceptance value. The information

exchanged among the workstations in the partition and the workstations reach to agreement about election of the leader. The acceptance value of each partition is then computed. The load balancing request is submitted to all the leaders in the network. The information is then exchanged among the leaders so as to reach to a consensus for allocating the load balancing request to the appropriate partition. Once the load is assigned to a given partition, members of that partition reach to agreement about accepting the jobs. The jobs that are not accepted by this partition are submitted again for load balancing.

3.1 Basics of Graph Theory

In this section we summarize some definitions and results from graph theory that will be useful in our analysis. A graph $G(V, E)$ consists of a number of vertices $V = \{V_1, V_2, \dots, V_k\}$ some of which are connected by edges $E(i, j)$ i.e. there is an edge between V_i and V_j . The vertices and edges may have non negative weights associated with it. The graph $G(V, E)$ is partitioned into l parts $P = \{P_1, P_2, \dots, P_l\}$ such that the size or weight of each partition is approximately the same i.e. $|P_i| = |P_j|$. In load balancing, each workstation of the network represents a vertex of the graph and each communication link connecting two workstations represents an edge between two vertices [7].

3.2 Splitting a Graph

A graph is said to be complete if each pair of its vertices is connected by an edge. A graph is said to be rigid if the distance between each pair of agents does not change over time under ideal conditions. Partitioning of graph means, splitting a rigid graph into two or more rigid parts. When splitting a rigid graph, it is necessary to break the links between points belonging to different partitions. However we preserve the links between two nodes belonging to the same partition (figures 3 and 4).

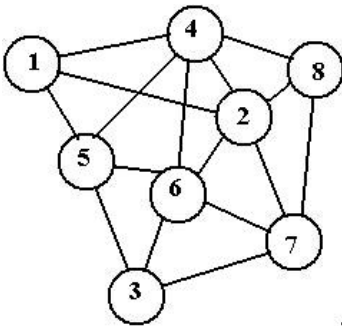


Fig. 3. Network before splitting

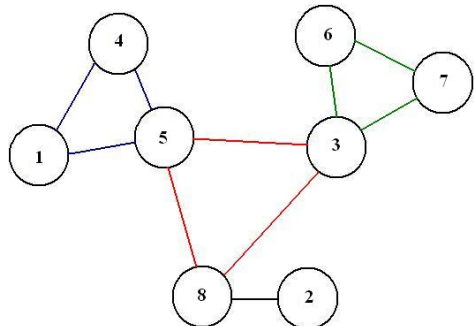


Fig. 4. Network after splitting

Let $G(V, E)$ be a rigid graph. Let V_1 and V_2 represent the two subsets of V such that $V_1 \cup V_2 = V$ and $V_1 \cap V_2 = \emptyset$. Let $E_1 \in E$ be the set of all edges whose both

end-vertices are in V_1 and $E_2 \in E$ be the set of all edges whose both end-vertices are in V_2 . Let $E_r = E \setminus (E_1 \cup E_2)$ be the set of all edges whose one end vertex is in E_1 and the other end vertex is in E_2 . Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$. When the graph $G(V, E)$ is split into $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, all edges in E_r are removed. The splitting problem is to find new sets of edges E_{1new} to insert into G_1 and E_{2new} to insert into G_2 such that the resulting graphs $\bar{G}_1 = (V_1, E_1 \cup E_{1new})$ and $\bar{G}_2 = (V_2, E_2 \cup E_{2new})$ are rigid.[8] Splitting a graph creates partitions which are balanced in terms of weight. Such partitions contain less number of workstations. Therefore state information needs to be exchanged among fewer workstations leading to less communication overhead. This approach could make the consensus based dynamic load balancing algorithm.

3.3 Information Exchange

In this section we discuss the information policy of the proposed consensus based dynamic load balancing algorithm. The aim of this information exchange is to define the actual state information that is to be exchanged among the workstations in the network and the individual partitions so as to achieve consensus.

In the distributed system, the inter- and intra-subsystem communication requirements form the basis for the overall system architecture. The workstations are connected by bidirectional communication links. It is therefore necessary for such workstations in a network to exchange information with each other so as to communicate with each other. Not only do such “information exchange” protocols enable the workstations to collect, disseminate, or update information, they also provide a basis to solve diverse problems in distributed computation.

The *information (I)* that is exchanged between the workstations is the acceptance value of each workstation and average acceptance value of the partition. The acceptance value is the function of balance memory space and CPU load of the workstation. $I = AV = f(c_i, m_i)$. The acceptance value can be calculated as follows [9]:

$$total_av(k) = \sum_{i=1}^n AV_k(i) \tag{1}$$

where $P_{mem}(t) = 1 - load_{mem}(t) / MAX_{mem}$. And MAX_{mem} is the maximum memory capacity, while $load_{mem}(t)$ is the actual memory load in MByte on the workstation, respectively. The memory parameter value is also time dependent and higher memory load on the workstation results in lower parameter value.

$P_{cpu}(t) = 1 - load_{cpu}(t) / MAX_{load}$ where $load_{cpu}(t)$ is the actual fraction (between 0 and 1) of the workstation’s CPU load. The CPU parameter value is time dependent and workstations with higher CPU load have lower value of this parameter. MAX_{load} is assumed to be unity. Consensus problems have a long history in computer science and form the foundation of the field of distributed computing. In this section consensus problem is explained in detail. The aim is to define the consensus equation which will decide which workstation(s) inside a given finally accepts the load balancing request.

3.4 Consensus

Consensus means, to reach an agreement regarding a certain quantity of interest that depends on the state of all agents in the network of agents or dynamic systems. The consensus problem in this work is to get a group of workstations in a network to agree upon a value in the presence of constraints. Consensus theory is used in the algorithm so that the leaders/workstations agree upon the acceptance value in order to distribute jobs uniformly among all the leaders/workstations such that none of the workstations is overloaded or under loaded. The jobs are assigned to that workstation inside a given partition whose acceptance value is greater than average acceptance value of that partition. [10] [11]

$AV_k(i)$ is the acceptance value of i^{th} workstation in the k^{th} partition, $av(k)$ is the average acceptance value of k^{th} partition and n is the total number of workstations in the k^{th} partition. $total_av(k)$ is the sum of the acceptance values of all the workstations in the k^{th} partition. Here the information variables $AV_k(i)$ and $av(k)$ represent an instance of consensus among the workstations in the network.

A set of n workstations is said to achieve consensus if

$$AV_k(i) - av(k) \geq 0 \text{ where } av(k) = (total_av(k)) / n \text{ and } total_av(k) = \sum_{i=1}^n AV_k(i)$$

The difference between the acceptance value of the individual workstation in a given partition and the average acceptance value of that partition should be greater than or equal to zero so that a given workstation is capable of receiving the jobs which constitutes the load balancing request.

3.5 Algorithm

1. Create the network and represent as graph with the nodes/workstations as vertices and connectivity as links.
2. Calculate the network weight by adding individual node weights of the workstations.
3. Calculate the partition weight (threshold value).
4. For (total number of partitions – 1)
 - a. Choose a pseudo-peripheral vertex as starting vertex of the partition
 - b. Partition weight = 0
 - c. **Repeat**
 - i. Partition weight = Partition weight + Node Weight
 - ii. If partition weight less than threshold value then add next adjacent connected node to this partition

Until the partition weight \leq threshold value
5. Preserving the links between vertices belonging to the same partition, remove other links
6. Add links, if necessary, to make partition minimally rigid
7. Calculate acceptance value of each workstation
8. Elect leader for each partition
9. Create links between the leaders.
10. Calculate acceptance value for partition using equation (1)
11. Repeat

- a. Submit the job to all leaders
- b. Choose the partition for accepting the batch of the jobs – Leaders bids for accepting and reach to consensus based on the acceptance value of the partition.
- c. **In the selected partition**
 - i. Leader submits the job in the partition
 - ii. Nodes bids for job and reach to consensus based on the acceptance value of the individual nodes.
 - iii. Update acceptance value of node and partition

Until all the jobs are distributed

12. Leader checks faulty workstations
 - a. If a workstation faulty, leader removes the jobs from this node and re-submits the job in the partition
 - b. Nodes bids for job and reach to consensus based on the acceptance value of the individual nodes
13. If new node joins the network, submit to leaders
 - a. Leaders bid for adding node and reach to consensus based on acceptance value.
14. If new load balancing request arrives, go to step 11
15. Stop

4 Simulation Results

In this section we present the simulation results of the proposed consensus based load balancing algorithm for different numbers of workstations, partitions and jobs. Also the comparison of proposed algorithm with important existing algorithm is given in Table 1.

Acceptance value of the workstation is calculated using two parameters: the CPU speed and the RAM size The CPU speed assumed varying from 2.1GHz to 3.2 GHz and the RAM size varying from 0.5 GB to 2 GB. The job is modeled using three parameters: CPU speed requirement, memory requirement and the expected computation time. It is assumed that the work stations are connected using high speed cables with no restriction on bandwidth. The algorithm is simulated for varying number of partitions, nodes and jobs. For simulation, the number of workstations is varied from

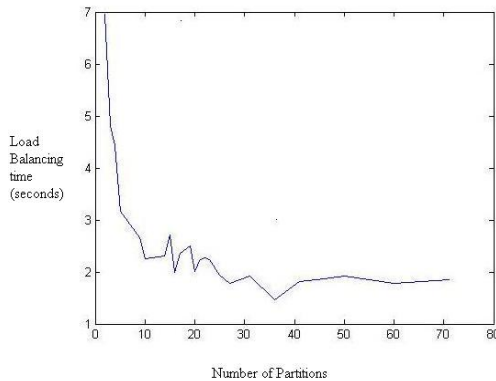


Fig. 5. Number of Partitions vs Load Balancing time

Table 1. Comparison of proposed algorithm with few existing algorithms

Sno	Metrics [12]	Consensus based Dynamic Load Balancing Algorithm	Bidding Approach [1]	Dynamic Scheduling Using Weights [5]	Dynamic Load Distribution Algorithm [6]
1	Scalability	Scalable up to a maximum limit of 1000 workstations	Maximum limit of 7 workstations	Maximum limit of 9 workstations	Maximum limit of 50 workstations
2	Approach	Dynamic decentralized	Fixed centralized	Fixed centralized	Fixed centralized
3	Number of jobs	Maximum number of Jobs = 2400 of 0.2 – 0.4 MB size	Maximum number of Jobs - 20 Size not specified	Job size and number of Jobs not specified	Job size and number of Jobs not specified
4	Load Balancing time	Load Balancing time varies from 0.18 to 400 secs	Load Balancing time upto 2.5 secs	Load Balancing time varies from 25 to 500 secs	Load Balancing time varies from 20 to 75 secs
5	Reliability & fault tolerance	Reliable & fault tolerant	Reliable & fault tolerant	Reliable & fault tolerant	Reliable & fault tolerant
6	Communication overhead	moderate	moderate	moderate	moderate

10 to 1000 while number of jobs varied from 10 to 4400. Figure 5 shows the plot of number of partitions vs load balancing time, Figure 6 shows plot of number of jobs vs load balancing time and plot of number of nodes load balancing time is shown in Figure 7. Load balancing time exponentially increases with increase in number of nodes as well number of jobs.

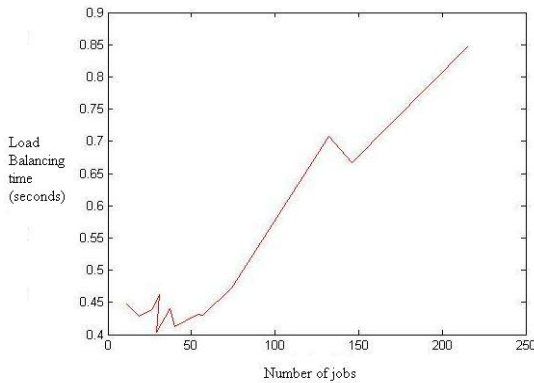


Fig. 6. Number of Jobs vs Load Balancing time

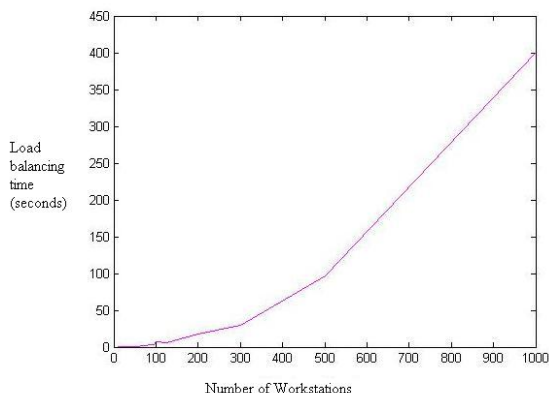


Fig. 7. Number of Workstations vs Load Balancing time

5 Conclusion

In this paper, we have proposed algorithm which is dynamic and decentralized. This algorithm is designed for a network of heterogeneous resources and is scalable up to a maximum limit of 1000 workstations. The job size varies between 0.1MB to 0.4 MB. For a partition containing 500 workstations, the maximum number of jobs that are accepted is approximately 4400. The algorithm is reliable and fault tolerant. Fault tolerance is incorporated by removing the faulty workstation. The jobs assigned to the faulty workstation are redistributed among the other workstations in that partition. The algorithm is reliable in the sense that a finite maximum number of jobs can be submitted to a given number of workstations inside a partition.

This algorithm maintains a balance between the load balancing time and communication overhead. The state information exchanged is the acceptance value and average acceptance value. The load balancing time increases as the number of workstations and the number of jobs increase (figures 5 and 6). For a given number of workstations the load balancing time decreases as the number of partitions increase (figure 7).

In the future stability of the proposed consensus based load balancing algorithm can be verified for variable job arrival rates. The algorithm can also be examined for scheduling interdependent jobs.

References

1. Janhavi, B., Surve, S., Prabhu, S.: Graph Partitioning for a Network of Heterogeneous Workstations. In: International Conference on Advances in Communication, Network and Computing, Kerala (2010)
2. Yagoubi, B., Slimani, Y.: Task Load Balancing Strategy for Grid Computing. *Journal of Computer Science* 3(3), 186–194 (2007)
3. Abed, Oz, G., Kostin, A.: Competition-Based Load Balancing for Distributed Systems. In: Proceedings of the Seventh IEEE International Symposium on Computer Networks, pp. 230–235 (2006)

4. Shah, R., Veeravalli, B., Misra, M.: On the Design of Adaptive and Decentralized Load-Balancing Algorithms with Load Estimation for Computational Grid Environments. *IEEE Transactions on Parallel and Distributed Systems* 18, 1675–1687 (2007)
5. Venu Gopalachari, M., Sammulal, P., Vinaya Babu Dr., A.: Correlating Scheduling and Load balancing to achieve optimal performance from a cluster. In: *International Advance Computing Conference*, pp. 320–325 (2009)
6. Chandra, P., Sahoo, B.: Dynamic Load Distribution Algorithm Performance in Heterogeneous Distributed System for I/O- intensive Task. In: *IEEE Region 10 Conference*, pp. 1–5 (2008)
7. Khan, M.S., Li, F.: Fast Graph Partitioning Algorithms. In: *Proc. IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, pp. 337–342 (1995)
8. Eren, T., Brian, D., Morse, A.S., Whiteley, W., Belhumeur, P.L.: Operations on Rigid Formations of Autonomous Agents. *Communications in Information and Systems* 3(4), 223–258 (2004)
9. Farcas, K., Hossman, T., Plattner, B., Ruf, L.: NWC: Node Weight Computation in MANETs'. In: *Proceedings of 16th International Conference on Computer Communications and Networks*, pp. 1059–1064 (2007)
10. Neiger, G.: Distributed Consensus Revisited. *Information Processing Letters*, 195–201 (1994)
11. Aspnes, J.: Randomized Protocols for Asynchronous Consensus. *Distributed Computing* 16(2-3), 165–175 (2003)
12. Janhavi, B., Surve, S., Prabhu, S.: Comparison of Load Balancing Algorithms in a Grid. In: *International Conference on Data Storage and Data Engineering, Bangalore*, pp. 20–23 (2010)

An Effective Way to Hide the Secret Audio File Using High Frequency Manipulation

Mahendra Kumar Pandey¹, Girish Parmar², and Sanjay Patsariya¹

¹ RJIT, BSF Academy, Tekanpur(M.P.)

mahendra2003in@yahoo.co.in,

sanjaypatsariya@gmail.com

² RTU, Kota(Raj.)

girish_parmar2002@yahoo.com

Abstract. As the technology increases day by day, human dependency increases on technologies. In present scenario everyone wants to do their work, business etc with the help of computer or with internet. Hence this increases the possibility of large-scale unauthorized copying which may lead to undermine the music, film, book and software industries. These concerns over protecting copyright have triggered significant research to find ways to hide copyright messages and serial number into digital media. An important sub discipline of information hiding is Watermarking. Digital Watermarking is an authentication technique which permanently embeds a digital signal (watermark) in text, image, audio, video files (any Data) by slightly modifying the data but in such a way that there are no harmful effects on the data. In this paper, we present audio watermarking technique that utilized the manipulation in high frequency of an audio signal to hide secret message The experimented result developed through MATLAB show that the usefulness of this technique.

Keywords: Audio watermarking, High frequency manipulation, Hamming window.

1 Introduction

Data hiding is a very active research subject of the areas of information security and multimedia signal. Due to development of science, now days human mostly depend upon computer in different area. To convey information from one place to another place, we need information security. To achieve this goal different method such as watermarking, cryptography, Steganography, coding etc. have been used. In recent year, among all method watermarking attract the people for information security.

Digital Watermarking is an authentication technique which permanently embeds a digital signal (watermark) in text, image, audio, video files etc. by slightly modifying the data in such a way that there are no harmful effects on the data. The watermark embedded may contain information such as identification of the product's owner, user's license information etc. This watermark can then be detected whenever required to identify its owner or to check whether a user is authentic to access that data or no. Fig.1. (a) show a generic diagram of digital watermarking.

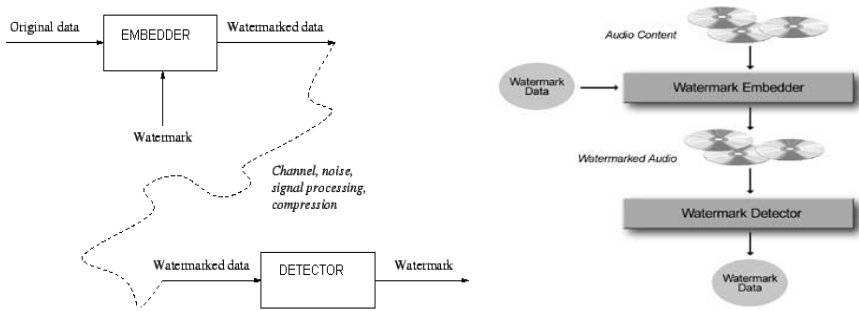


Fig. 1. a) A generic diagram of watermarking. (b) A schematic diagram of audio watermarking

Audio watermarking is a technology which allows a secret message (watermark) to be hidden in an audio file while preserving the integrity of the original file. As shown in fig 1. (b), The watermark is embedded such that it is not affected by any processing of the watermarked audio signal. The watermark is embedded in such parts of the signal if any one tries to remove it, the quality of the signal does not affected. The watermark is not apparent to the user; watermark information is predominantly used to identify the creator or the authentic person of an audio file. [4]

2 Trade-Offs Exist between the Quantity of Data and the Immunity to Modification

Trade-offs exists between the quantity of embedded data and the degree of immunity to host signal modification. By constraining the degree of host signal degradation, a data-hiding method can operate with either high added data rate, or high resistance to modification, but not both. As one increases, the other must decrease. In any system, we can trade bandwidth for robustness by exploiting redundancy.

The quantity of embedded data and the degree of host signal modification vary from application to application. Consequently, different techniques are employed for different applications [8].

3 Properties of Audio Watermarks

3.1 Inaudibility:-The watermark embedded into the audio signal should not degrade the sound quality.

3.2 Robustness:-The watermark should resist any transformations applied to the audio signal as long as the sound quality is not unacceptably degraded.

3.3 Capacity:-The watermark bit rate must be high enough for the intended application, which can be conflicting with in audibility and robustness.

3.4 Reliability:-The data contained in the watermark should be extracted with acceptable error rates.

3.5 Low Complexity:-When the watermark is used for real-time applications, watermarking algorithm should not be excessively time-consuming.

4 Method of Watermarking

There are a large number of techniques in digital audio watermarking namely Echo Watermarking Amplitude Variation, spread spectrum, here we will present the new approach.

4.1 High Frequency Manipulation Audio Watermarking

This technique relies on basic principle that on varying the high frequency of the host signal by a few percent produces inaudible variations resulting in a watermarked signal which sounds identical to the original host file. To implement this idea we take the advantage of human auditory system, we know that the frequency range of human hearing only extends from 15 Hz up to somewhere between 15000 and 20000 Hz depending on the individual. Even those who can hear up to 20000 Hz cannot hear those very high frequencies well. So there is the potential or key of our approach to alter the high frequency range of a sound by inserting a secret message and have the result be imperceptible.

4.2 Watermark Embedding

Fig 3. show basic watermark embedding process which can be represented as

$$y(n) = x(n) + w(n) \quad (1)$$

Where $y(n)$ watermarked audio signal, $x(n)$ the original signal, and $w(n)$ the watermark signal. [1]

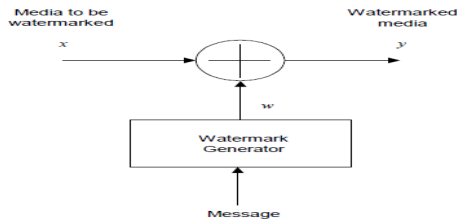


Fig. 3. Basic watermark embedding process

In this method, we embed data into an original audio signal. This encoding process involves 3 essential steps:-

- Preparing the carrier signal
- Preparing the secret message
- Adding the signals together

The resulting signal is then scaled to avoid clipping. Finally we get the composite signal.

4.3 Preparing the Carrier Signal

For preparing the carrier signal, first of all we have to select carrier sound such that it should be long enough to fit the entire secret message. carrier signal should be a relatively active signal that contains a large amount of information in the low and mid range frequencies, so that we can easily insert our secret message on high frequency part of this carrier signal. Now we passes carrier through a low pass filter with a cutoff of 17000 Hz to cordon off the 17000-22050 Hz frequency range for our secret message.

we consider the ideal, or "brick wall," digital low pass filter with a cutoff frequency of ω_0 rad/s. This filter has magnitude 1 at all frequencies with magnitude less than ω_0 , and magnitude 0 at frequencies with magnitude between ω_0 and π . Its impulse response sequence $h(n)$ is given by this equation.

$$h(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(\omega) e^{j\omega n} d\omega = \frac{1}{2\pi} \int_{-\omega_0}^{\omega_0} e^{j\omega n} d\omega = \frac{\omega_0}{\pi} \text{sinc}\left(\frac{\omega_0}{\pi} n\right) \quad (2)$$

This filter is not implementable since its impulse response is infinite and non causal. To create a finite-duration impulse response, truncate it by applying a window. By retaining the central section of impulse response in this truncation, you obtain a linear phase FIR filter. For example we take, a length 51 low pass filter with a cutoff frequency ω_0 of 0.4π rad/s is

$$b = 0.4 * \text{sinc}(0.4 * (-25:25));$$

The window applied here is a simple rectangular window. By Parseval's theorem, this is the length 51 filter that best approximates the ideal low pass filter, in the integrated least squares sense. [4]

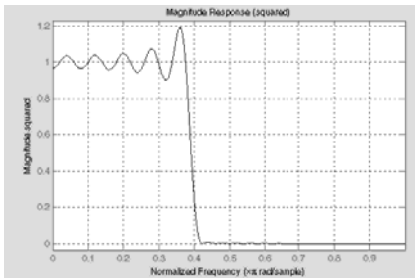


Fig. 4. (a)

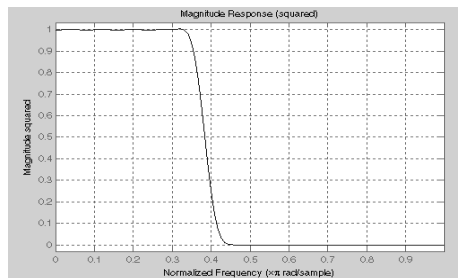


Fig. 4. (b)

As shown in Fig. 4. (a), ringing and ripples occur in the response, especially near the band edge. This "Gibbs effect" does not vanish as the filter length increases, but a nonrectangular window reduces its magnitude. Multiplication by a window in the time domain causes a convolution or smoothing in the frequency domain. Apply a length 51 Hamming window to the filter and this reduces the ringing effect as shown in Fig. 4(b).

$$b = 0.4 * \text{sinc}(0.4 * (-25:25));$$

$$b = b.*\text{hamming}(51)$$

This improvement is at the expense of transition width (the windowed version takes longer to ramp from pass band to stop band) and optimality (the windowed version does not minimize the integrated squared error).

5 Preparing the Secret Message

The exact method of preparation for the secret message depends on its nature (sound, image, etc.), but we essentially formatting the data so that it fits roughly within a 17000-22000 Hz frequency range. For this purpose we pass the secret sounds through a band pass filter with a 4000 Hz window. And multiply the message by a 20 KHz cosine wave to modulate it so that the frequency range of the encoded secret message is pushed up 20 kHz. Finally , when Once both the carrier and the secret message have been prepared, they can be simply added together to get the encoded signal. The resulting signal is then scaled to avoid clipping. Finally we get the composite signal.

6 Decoding Process

For decoding the secret sound file, we have to pass the composite signal through a Band pass filter having the band of the modulation freq (20000 Hz) +/- the max secret freq so we can easily remove the carrier signal with composite signal. After getting this, Once again we multiply the message by a 20 kHz cosine wave to the secret message back to the correct frequency range. Now using Low pass filter with cutoff frequency of 3300 , we pass this secret signal to get rid of any extra chatter. Again the resulting signal is then scaled to avoid clipping.

7 Results

All results shown in Fig. 5. Sample1 and Sample2 have been obtained through MATLAB a Using high frequency manipulation technique.

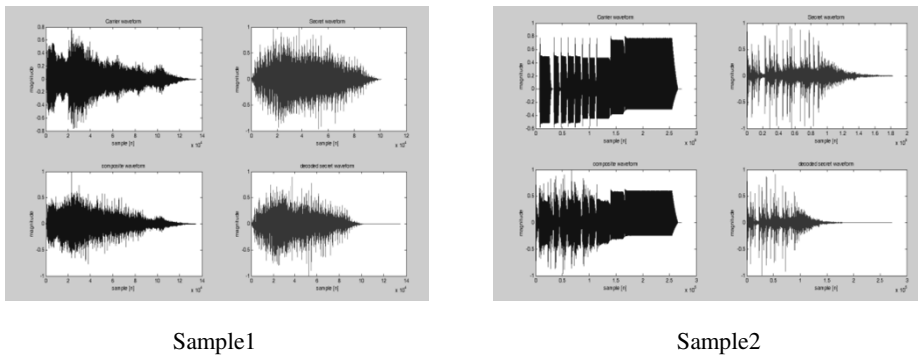


Fig. 5.

From results as shown in fig 5, we say that performance of this algorithm is clearly dependent on human analysis and judgment. This High frequency manipulation techniques has shown capacity to protect audio files over threat of due to its ease of implementation and robustness.

8 Conclusion

We proposed an audio watermarking technique that exploits the fact that the perception of the auditory system for stereo audio images is relatively immune in the high-frequency regions. Results show that secret message is successfully embedded with help of algorithm developed in MATLAB. It is easy to implement, self sufficient blind watermarking algorithm .through this paper we shows the system is robust against perceptual audio coding, reverberations and additive Gaussian noise. Detecting the existence of the watermarking is difficult without knowing the key used in the encoding process.

References

- [1] Katzenbeisser, S., Petitcolas, F.A.P.: Information Hiding: Techniques for steganography and digital watermarking. Artech House, Boston (1999)
- [2] Lie, W.-N., Chang, L.-C.: Robust and High-Quality Time-Domain Audio Watermarking Based on Low-Frequency Amplitude Modification. *IEEE Transaction on Multimedia* 8(1), 48–52 (2006)
- [3] Pholsomboon, S., Vongpradhip, S.: Rotation, Scale, and translation Resilient Digital Watermark Based on Complex Exponential Function. *ECTI Transactions on Electric Eng., Electronics and Communication* 2(2), 40–47 (2004)
- [4] Wikipedia /Internet, <http://en.wikipedia.org/wiki/Encryption>, <http://mathworks.com>
- [5] Larbi, S.D., Jaïdane-Saïdane, M.: Audio Watermarking: A Way to Stationnarize Audio Signals. *IEEE Transaction on Signal Processing* 53(2), 816–822 (2005)
- [6] Dugelay, J., Roche, S.: A Survey of Current Watermarking Techniques. In: Katzenbeisser, S.C., et al. (eds.) *Information Techniques for Steganography and Digital Watermarking*, pp. 121–145. Artec House, Northwood (December 1999)
- [7] Kutter, M.: *Digital Watermarking Frequently Asked Questions* (2001), <http://www.watermarkingworld.org/faq.html>currentMarch2004
- [8] Bender, W., Gruhl, D., Morimoto, N., Lu, A.: Techniques for data hiding. *IBM Systems Journal* 35, 313–336 (1996)

Web Usage Mining: An Implementation View

Sathya Babu Korra, Saroj Kumar Panigrahy, and Sanjay Kumar Jena

Department of Computer Science and Engineering
National Institute of Technology Rourkela, 769 008, Odisha, India
{ksathyababu, panigrahys, skjena}@nitrkl.ac.in

Abstract. This paper describes the implementation of Web usage mining for DSpace server of NIT Rourkela. The DSpace log files have been preprocessed to convert the data stored in them into a structured format. Thereafter, the general procedures for bot-removal and session-identification from a Web log file have been applied with certain modifications pertaining to the DSpace log files. Furthermore, analysis of these log files using a subjective interpretation of recently proposed algorithm EIN-WUM has also been conducted.

Keywords: Data mining, Web data, Web usage mining, DSpace.

1 Introduction

Web usage mining (WUM) is the process of applying data mining techniques to the discovery of usage patterns from Web data, targeted towards various applications [1]. WUM involves mining the usage characteristics of the users of Web applications. This extracted information can then be used in a variety of ways such as— improvement of the application, checking of fraudulent elements etc. The major problem with Web mining in general and WUM in particular is the nature of the data they deal with. With the upsurge of Internet in this millennium, the Web Data has become huge in nature and a lot of transactions and usages are taking place by the seconds. Apart from the volume of the data, the data is not completely structured. It is in a semi-structured format and needs a lot of preprocessing and parsing before the actual extraction of the required information. This paper describes about the work in which, a small part of the WUM process has been taken up, that involves preprocessing, user identification, bot-removal and analysis of the log files of DSpace Web server at NIT Rourkela.

2 Data for Web Usage Mining

In Web Mining, data can be collected at the server-side, client-side, proxy servers, or obtained from an organization's database (which contains business data or consolidated Web data for business intelligence [2]). Each type of data collection differs not only in terms of the location of the data source, but also the kinds of data available, the segment of population from which the data was collected, and its method of implementation.

Web Data: The various kinds of data that can be used in Web mining are *Content*— usually consists of multimedia contents such as text, graphics, etc; *Structure*— describes the organization of the contents, i.e., HTML or XML tags, hyperlinks, etc.; *Usage*— the pattern of usage of webpages such as IPs, page references, and the date and time of access; and *User Profile*— demographic information about users of the website such as registration data and customer profile information [1].

Data Sources: The data sources may include Web data repositories— *Web Server Logs*, i.e., a history of page requests [3,4]; *Proxy Server Logs*, i.e., proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers and serve as a data source for characterizing the browsing behavior of a group of anonymous users, sharing a common proxy server; *Browser Logs*, i.e., client-side data collection can be done by using a remote agent (such as JavaScript or Java applets) or by modifying the source code of an existing browser (such as Mozilla) to enhance its data collection capabilities [1].

Abstract Data: The information obtained by the data sources described above can be used to identify various abstract data— number of hits, number of visitors, visitor referring website, visitor referral website, time and duration, path analysis, browser type, cookies, and platform [5].

Possible Actions: The data collected can be analysed and the following possible actions can be taken— shortening paths of high visit pages, eliminating or combining low visit pages, redesigning pages to help user navigation, and helping in evaluating effectiveness of advertising campaigns [5].

3 Web Usage Mining

There are three main tasks for performing WUM— preprocessing, pattern discovery and pattern analysis [1]. These are briefly explained as follows.

Preprocessing: Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. The different types of preprocessing in WUM are— *usage*, *content*, and *structure* preprocessing.

Pattern Discovery: WUM can be used to uncover patterns in server logs but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data. The various pattern discovery methods are— *Statistical Analysis*, *Association Rules*, *Clustering*, *Classification*, *Sequential Patterns*, and *Dependency Modeling*.

Pattern Analysis: The need behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The most

common form of pattern analysis consists of a knowledge query mechanism such as SQL. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

4 Implementation Details

This section describes the various operations that have been done for finding web usage patterns of DSpace server of NIT Rourkela. Different web server log analyzers like Web Expert Lite 6.1 and Analog 6.0 have been used to analyze various sample web server logs obtained. The key information obtained was—total hits, visitor hits, average hits per day, average hits per visitor, failed requests, page views total page views, average page views per day, average page views per visitor, visitors total visitors average visitors per day, total unique IPs, bandwidth, total bandwidth, visitor bandwidth, average bandwidth per day, average bandwidth per hit, average bandwidth per visitor; access data like files, images, referrers, user agents etc.

4.1 Collection of DSpace Log Files

The DSpace server log files were collected and the features found are shown below. The Common log contains the requested resource and a few other pieces of information, but does not contain referral, user agent, or cookie information. The information is contained in a single file. The following example shows these fields populated with values in a common log file record:

host	log	user	date:time GMToffset	request	status	bytes
125.125.125.125	-	-	[10/Oct/1999:21:15:05 +0500]	"GET /index.html HTTP/1.0"	200	1043

4.2 Analysis of Web Server Logs

First part of analysis was preprocessing. Preprocessing segregated all the details provided in the log file into a structured form. JAVA is used for this. Data structures used are linear arrays—ip, time, content, httpmethod, httpstatus, bandwidth, browser etc.

4.3 Key Constraints and Solutions

Not much Variation in IP: As we are considering the DSapce log files, which are specific to NIT Rourkela, it is observed that there is not much variation in IP addresses in the entries recorded in the log file.

Usernames and Aliases not Provided: The second and third entries in the common log format are the usernames and aliases which are mainly recorded in a login based website. These information are not there in the DSpace log files.

Web Crawlers: The various types of crawlers found in the DSpace log files are— MSN bots, Yahoo slurps, Google bots, Baidu spiders etc.

Bot Identification: After much analysis of bot identification and removal, a method has been used specific to DSpace log files to do the same. The pseudocode for the method is as follows:

```

BotId()
{ while(!EOF)
  { readLine();
    Check for keywords (bot,slurp,spider) in browser[] array
    if the array contains keyword
      { botflag=true;  botcounter++; }
    else
      botflag=false;
  } }

```

Identification of User Sessions: User sessions in WUM generally refers to the usage or access of any content of the website from a fixed IP over a fixed period of time. The period of time is subjective to the analyzer. Considering the above requirements, a method specific to DSpace log file has been used to identify user sessions in the log file. The pseudocode for the method is as follows:

```

SessionId()
{ while(!EOF)
  { i=1;
    add first not bot entry to session i;
    for each (next entry)
      { if(entry != bot)
        if(IP == previous IP)
          if(time[this entry] - time[this entry -1] < x)
            add entry to session i;
          else
            { i++; add entry to session i; }
        else
          { i++; add entry to session i; }
      } } }

```

4.4 Using EIN-WUM Algorithm

After preprocessing, bot identification and removal, and session identification, the EIN-WUM (Enhanced Immune Network Web Usage Mining) algorithm [6] is used. Our interpretation of the algorithm subject to DSpace website of NIT Rourkela is as follows:

- Limit value of no. of antibodies to 6 (based on the category from DSpace Website).
- We define the category of each entry in the Server Log by assigning it a number (0 through 6). The numbers signify—0 - default value, 1 - content

searched by title, 2 - content searched by author, 3 - content searched by date, c - Content searched by author, 5 - content accessed by handle, and 6 - content accessed by bitstream.

- The antibodies are initialized from the first 10 sessions. For each session an entry goes to the corresponding number of antibody as its category is. So each antibody contains only one category of server log entry.
- For each incoming session, compare with each existing antibody. If (similarity of antibody > threshold), replace old session with new session, else if (similarity < threshold) update antibody with most similarity.
- Put a limit on the size of antibody. If (antibody crosses limit), delete old entries.

The various Utilities of the above interpretation are found as:

- a. At the end of the program, the ten most interesting antibodies will remain.
- b. The contents accessed in the antibodies will be the most frequently accessed contents in the whole website.
- c. Based on (b) the following changes can be brought to the concerned site:
 - i. improvements on frequently accessed pages.
 - ii. deletion or merging of unused pages.
 - iii. improvement of content.
 - iv. improvement of interaction with referral sites.

4.5 Results

The results obtained from the analysis are given below.

Preprocessed Information from Log Files: The preprocessing program collected the details in the appropriate data structures and also identified whether an entry is a bot entry or a valid user entry (shown below).

1	true	203.129.199.129	10/Jan/2010:04:04:26	GET	200
	17013B	0	/dspace/browse-title?top=2080%2F905		
2	true	203.129.199.129	10/Jan/2010:04:04:29	GET	200
	14295B	0	/dspace/browse-author?bottom=Misra%2C+M		

Summary of the Log File and Sessions: The summary of the log file giving overall details, the sessions and the different log file entries that constitute the sessions are shown below.

*****Summary*****	session 1	182
number of hits = 14274	session 1	183
number of visitor hits= 7923	session 1	191
number of spider hits= 6351	session 2	193
Number of days= 5	session 2	194
Average hits per day = 2854	session 2	195
Total Bandwidth used = 1494419892 Bytes	session 2	196
Avenrage Bandwidth= 298883978 Bytes	session 2	197
*****	session 2	198

Usage Patterns: The different frequently accessed contents in the DSpace website is shown below.

```

16  0  1  /dSPACE/browse-author?starts_with=Das%2C+Atanau
92  0  1  /dSPACE/browse-author?bottom=Wai%2C+P+K+A
181 0  1  /dSPACE/browse-author?starts_with=Verghese%2C+L
227 1  1  /dSPACE/browse-author?top=Joshi%2C+Avinash
364 1  1  /dSPACE/browse-author?top=Joshi%2C+Avinash
527 5  1  /dSPACE/browse-author
530 5  1  /dSPACE/browse-author?starts_with=C
532 5  1  /dSPACE/browse-author?top=Chatterjee%2C+Saurav
536 5  1  /dSPACE/browse-author?starts_with=S
569 7  1  /dSPACE/browse-author?starts_with=S
571 7  1  /dSPACE/browse-author?top=Chatterjee%2C+Saurav
715 8  1  /dSPACE/browse-author?top=Bal%2C+S
748 8  1  /dSPACE/browse-author?starts_with=Das%2C+B+M
831 8  1  /dSPACE/browse-author?starts_with=Karanam%2C+U+M+R

```

5 Conclusions

The proposed methods were successfully tested on the log files for bot removal and user sessions identification. The results which were obtained after the analysis were satisfactory and contained valuable information about the log files. The methodology and implementation presented in this paper are purely DSpace Website specific. Analysis of above obtained information proved WUM as a powerful technique in Website management and improvement. However, this subjective interpretation of the algorithm EIN-WUM is very ingenious and proposes a lot of scope to be extended on to other problem domains.

References

1. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explor. Newsl. 1(2), 12–23 (2000), <http://portal.acm.org/citation.cfm?id=846188>
2. Abraham, A.: Business intelligence from web usage mining. Journal of Information & Knowledge Management, iKMS & World Scientific Publishing Co. 2(4), 375–390 (2003), <http://www.worldscinet.com/jikm/02/0204/S0219649203000565.html>
3. W3C: Logging control in w3c httpd, <http://www.w3.org/Daemon/User/Config/Logging.html>
4. W3C: Extended log file format. W3c working draft wd-logfile-960323, <http://www.w3.org/TR/WDlogfile.html>
5. Gupta, G.K.: Introduction to Data Mining with Case Studies. Phi Learning, 1st edn. (2008)
6. Rahmani, A.T., Helmi, B.H.: Ein-WUM: an AIS-based algorithm for web usage mining. In: Ryan, C., Keijzer, M. (eds.) GECCO, pp. 291–292. ACM, New York (2008), <http://doi.acm.org/10.1145/1389095.1389144>

A Genetic Algorithm Way of Solving RWA Problem in All Optical WDM Networks

Ravi Sankar Barpanda*, Ashok Kumar Turuk,
Bibhudatta Sahoo, and Banshidhar Majhi

Department of Computer Science & Engineering
National Institute of Technology Rourkela Orissa India
{barpandar, akturuk, bdsahu, bmajhi}@nitrkl.ac.in
<http://www.nitrkl.ac.in>

Abstract. This research work presents a Genetic Algorithm (GA) heuristic approach to solve the static Routing and Wavelength Assignment (RWA) problem in Wavelength Division Multiplexing (WDM) networks under wavelength continuity constraint. The RWA problem is modeled as an Integer Linear Programming (ILP) problem with the optimization objective to balance the network load among the connection requests. We consider ARPANET as the standard simulation network and use Genetic Algorithm technique to solve the formulated ILP on such network to produce a near optimal solution in polynomial time. We state three different fitness functions, all of them aim at balancing the network load among individuals and compare them while optimizing different network parameters.

Keywords: Genetic Algorithm; RWA problem; WDM network; wavelength continuity constraint; Integer Linear Programming; fitness function.

1 Introduction

Routing and Wavelength Assignment (RWA) [1] is a well known issue in Wavelength Division Multiplexing (WDM) optical networks [2]. In such networks, each fiber link is logically divided into multiple number of non interfering wavelength channels. The RWA problem assumes determining the routes and wavelengths to be used to create the lightpaths [1] for the connection requests. The RWA problem can be separated into two sub-problems, routing allocation and wavelength channel assignment. The first subproblem determines the physical links that will define each lightpath while the second subproblem assigns wavelength(s) to each lightpath. The RWA problem has been previously considered for various design objectives, for instance, minimizing the number of wavelengths required on each edge while satisfying all connection requests in the demand matrix or maximizing the number of accepted connection requests given a limited number of wavelength channels per fiber link. The optimal solution to the RWA problem is found to be NP-hard [3] and thus suited to heuristic methods [4-6].

* This research is supported by SERC, DST, Government of India and monitored by Centre for Soft computing Research: A National Facility, ISI, Kolkata.

2 Proposed Work

We consider the static version of the RWA problem and optimize different RWA objective criteria enumerated as follows:

- minimizing the congestion of the most congested link in the network
- minimizing the difference between most congested and least congested link
- minimizing the difference between most congested link and average congestion of all links in the network.

The Min-RWA problem is modeled as an Integer Linear Programming (ILP) problem with suitable constraints to establish loop free resilient lightpaths. The formulated ILP tailored with Genetic Algorithm (GA) heuristic is implemented on ARPANET (Advanced Research Project Agency NETwork) to produce a near optimal solution.

3 Statement of the Problem Formulation

We assume a network that maintains wavelength continuity constraint. The given optical network is supposed to be single fiber. The optical network is viewed as a graph $G = (V, E)$ where V is the set of nodes and E is the set of undirected edges. Let W be the set of wavelengths supported by every fiber link in the network and K be the set of static lightpath requests. The demand matrix is specified by D where D_{ij} defines the maximum demand between node pair i and j . The variables of concern of the formulated ILP are defined as follows:

$$x_k^w = \begin{cases} 1; & \text{if the lightpath } k \text{ is established with wavelength } w \\ 0; & \text{otherwise} \end{cases}$$

$$x_k^{w,e} = \begin{cases} 1; & \text{if the lightpath } k \text{ is established with wavelength } w \text{ on link } e \\ 0; & \text{otherwise} \end{cases}$$

The network design formulations stated here is to optimize three different objective functions:

$$\text{Minimize } \max_{e \in E} \sum_{k \in K} \sum_{w \in W} x_k^{w,e} \quad (1)$$

$$\text{Minimize } \max_{e \in E} \sum_{k \in K} \sum_{w \in W} x_k^{w,e} - \min_{e \in E} \sum_{k \in K} \sum_{w \in W} x_k^{w,e} \quad (2)$$

$$\text{Minimize } \max_{e \in E} \sum_{k \in K} \sum_{w \in W} x_k^{w,e} - \frac{\sum_{k \in K} \sum_{e \in E} \sum_{w \in W} x_k^{w,e}}{|E|} \quad (3)$$

subject to:

– Wavelength continuity constraint:

$$\sum_{w \in W} x_k^w \leq 1; \forall k \in K \quad (4)$$

– Wavelength distinct constraint:

$$\sum_{k \in K} x_k^{w,e} \leq 1; \forall w \in W \text{ and } \forall e \in E \quad (5)$$

– Demand constraint:

$$\begin{aligned} & \{ \{k \in K \mid \sum_{e \in \omega^-(i)} \sum_{w \in W} x_k^{w,e} - \sum_{e \in \omega^+(i)} \sum_{w \in W} x_k^{w,e} = -1 \\ & \wedge \sum_{e \in \omega^+(j)} \sum_{w \in W} x_k^{w,e} - \sum_{e \in \omega^-(j)} \sum_{w \in W} x_k^{w,e} = -1 \} \} \leq D_{ij} \end{aligned} \quad (6)$$

– Wavelength reservation constraint:

$$\sum_{e \in \omega^-(v): v \in V - \{s_k, d_k\}} x_k^{w,e} - \sum_{e \in \omega^+(v): v \in V - \{s_k, d_k\}} x_k^{w,e} = 0; \forall k \in K \text{ and } \forall w \in W \quad (7)$$

– No looping constraint around source node(s_k):

$$\sum_{e \in \omega^-(s_k): s_k \in V} \sum_{w \in W} x_k^{w,e} = 0; \forall k \in K \quad (8)$$

– No looping constraint around destination node(d_k):

$$\sum_{e \in \omega^+(d_k): d_k \in V} \sum_{w \in W} x_k^{w,e} = 0; \forall k \in K \quad (9)$$

– No looping constraint around intermediate nodes:

$$\sum_{e \in \omega^-(v): v \in V - \{s_k, d_k\}} \sum_{w \in W} x_k^{w,e} \leq 1; \forall k \in K \quad (10)$$

$$\sum_{e \in \omega^+(v): v \in V - \{s_k, d_k\}} \sum_{w \in W} x_k^{w,e} \leq 1; \forall k \in K \quad (11)$$

$$\sum_{e \in \omega^+(v): v \in V - \{s_k, d_k\}} \sum_{w \in W} x_k^{w,e} - \sum_{e \in \omega^-(v): v \in V - \{s_k, d_k\}} \sum_{w \in W} x_k^{w,e} = 0; \forall k \in K \quad (12)$$

– Hop-Count Constraint:

$$\sum_{e \in E} \sum_{w \in W} x_k^{w,e} \leq H \text{ where } H = \max_{(s_k, d_k)} \{d(s_k, d_k)\} + \alpha \quad (13)$$

where $d(s_k, d_k)$ is the minimum distance between a node pair (s_k, d_k) and the parameter α depends on the routing heuristic.

4 The GA Approach to Solve the RWA Problem

4.1 The Chromosome Structure

The chromosome is a group of vectors coded as $\begin{bmatrix} p_1 \\ \vdots \\ p_{|K|} \end{bmatrix}$ where each vector p_i is a lightpath represented as $(n_{i0} \dots n_{ih(i)})$; $n_{i0}, \dots, n_{ih(i)} \in V$.

4.2 Initial Population

For every lightpath (s_i, d_i) , employ Dijkstra's algorithm to find the minimum cost path p_i . All the p_i 's form the first chromosome of the first iteration. For each fiber link $(n_{ij}, n_{i(j+1)})$ in every p_i , disable one link at a time and find minimum cost paths to form a new chromosome. Repeat the process until the population size is reached. In this work, a population size of 50 is maintained.

4.3 Fitness Function

We state three different fitness functions corresponding to different RWA objective criteria as stated in Eq. 1, Eq. 2 and Eq. 3.

$$\begin{aligned} y_1 &= 1 - \frac{\text{max_con}}{|K|} \\ y_2 &= 1 - \frac{\text{max_con} - \text{min_con}}{|K|} \\ y_3 &= 1 - \frac{\text{max_con} - \text{avg_con}}{|K|} \end{aligned} \quad (14)$$

where

max_con = congestion of the most congested link in the network

min_con = congestion of the least congested link in the network

avg_con = average congestion of all links in the network

$|K|$ = number of static lightpaths

4.4 Selection of Chromosomes for the Next Generation

The chromosomes of the next generation are selected from the current population by a spinning roulette wheel method [7].

4.5 Crossover

In the selected generation, with a certain crossover probability a chromosome is asked for mating with another chromosome. According to a crossover ratio, calculate the number of lightpaths that will be modified. Pick these lightpaths randomly and exchange the picked lightpaths between the two chromosomes. In the simulation work, the crossover rate is maintained at 0.5 and the crossover ratio is limited to 0.2.

Table 1. Demand set of static lightpath requests

Lightpath(s_i, d_i)	Minimum cost path(p_i)
n10-n13	n10-n19-n20-n17-n13
n15-n7	n15-n14-n20-n19-n10-n7
n7-n5	n7-n3-n1-n2-n5
n10-n1	n10-n7-n3-n1
n15-n14	n15-n14
n10-n8	n10-n8
n8-n15	n8-n12-n14-n15
n2-n10	n2-n1-n3-n7-n10
n3-n9	n3-n8-n9
n9-n20	n9-n10-n19-n20

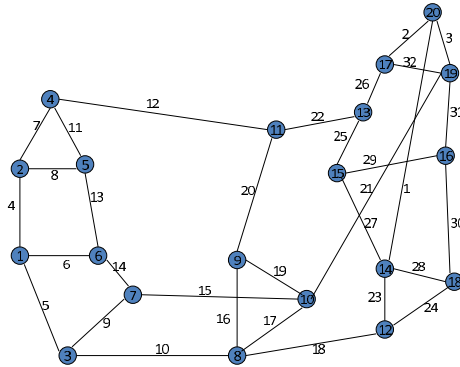


Fig. 1. Advanced Research Project Agency Network

4.6 Mutation

In the selected generation, with a certain mutation rate a chromosome is mutated. According to a mutation ratio, calculate the number of lightpaths that will be modified. For each such lightpath $p_i = [n_{i0}, n_{i1}, \dots, n_{ih_i}]$, randomly pick two adjacent nodes n_{ij} and $n_{i(j+1)}$. Disable the fiber link between the two nodes and remove all the nodes n_{il} such that $l < j$ and $l > j + 1$. Calculate the minimum cost path between n_{ij} and $n_{i(j+1)}$; replace the corresponding portion in p_i using the new path. In the simulation work, the mutation rate is maintained at 0.1 and the mutation ratio is limited to 0.2.

5 Simulation Result

The formulated ILP tailored with GA is simulated on ARPANET (see Fig. 1) to satisfy a set of static lightpath requests which is listed in Tab. 1. The stopping criterion for the GA is the maximum number of generations which is restricted to 100 in this simulation work. The performance comparison among the RWA objectives is based on their ability to optimize various network parameters and is shown in Tab. 2.

Table 2. Performance comparison among RWA objective criteria

Type of fitness functions	comparison
fitness function:y ₁	The maximum fitness of a chromosome after required number of generations is: 0.800 The network load for establishing 10 lightpaths is: 02 The total delay in establishing all the lightpaths is: 505 The maximum hops traversed by a lightpath: 07 The number of fibers used to honor all the lightpaths: 24 The maximum delay of a lightpath: 100
fitness function:y ₂	The maximum fitness of a chromosome after required number of generations is: 0.9 The network load for establishing 10 lightpaths is: 02 The total delay in establishing all the lightpaths is: 433 The maximum hops traversed by a lightpath: 06 The number of fibers used to honor all the lightpaths: 22 The maximum delay of a lightpath: 68
fitness function:y ₃	The maximum fitness of a chromosome after required number of generations is: 0.956 The network load for establishing 10 lightpaths is: 02 The total delay in establishing all the lightpaths is: 793 The maximum hops traversed by a lightpath: 10 The number of fibers used to honor all the lightpaths: 29 The maximum delay of a lightpath: 178

6 Conclusion and Future Work

Among the stated RWA objectives, the objective of minimizing the difference between most congested and least congested link in the network provides best performance while optimizing different network parameters such as congestion, delay, maximum hop count and total number of fibers used to honor all the lightpaths. The simulation work may be extended further to accommodate different sets of lightpath requests and the ability of the RWA objectives can be compared as they optimize various network parameters while establishing these sets of lightpath requests; thereby analyzing any alteration in their performances.

References

1. Zang, H., Jue, J.P., Mukherjee, B.: A Review of Routing and Wavelength Assignment Approaches for Wavelength Routed Optical WDM Networks. *Optical Networks Magazine* 1(1), 47–60 (2000)
2. Mukherjee, B.: *Optical WDM Networks*. Springer, Heidelberg (January 2006)
3. Banerjee, D., Mukherjee, B.: A practical Approach for Routing and Wavelength Assignment in Large Wavelength-Routed Optical Networks. *IEEE Journal on Selected Areas in Communications* 14(5), 903–908 (1996)
4. Qin, H., Liu, Z., Zhang, S., Wen, A.: Routing and Wavelength Assignment based on Genetic Algorithm. *IEEE Communication Letters* 6(10), 455–457 (2002)
5. Sinclair, M.C.: Minimum Cost Wavelength-Path Routing and Wavelength Allocation Using a Genetic Algorithm/ Heuristic hybrid Approach. *IEEE Proceedings on Communications* 146(1), 1–7 (1999)
6. Saha, D., Purkayastha, M.D., Mukherjee, A.: An Approach to Wide area WDM Optical Network Design Using Genetic Algorithm. *Computer Communications* 22, 156–172 (1999)
7. Pan, Z.: Genetic Algorithm for Routing and Wavelength Assignment Problem in All-optical Networks. Technical report, Department of Electrical and Computer Engineering, University of California, Davis (2002)

Working of Web Services Using BPEL Workflow in SOA

Aarti M. Karande, Vaibhav N. Chunekar, and B.B. Meshram

V.J.T.I., Matunga, Mumbai 400019

aartimkarande@gmail.com, chunekar@rediff.com,

bbmeshram@vjti.ac.in

Abstract. Service Oriented Architecture is used to achieve loose coupling among diverse interacting software applications. SOA is used for reduction in development time and cost. Web services standards used for SOA are distributed software components that provide information to applications rather than to humans, through an application-oriented interface. SOA with web services standards provide greater interoperability. It also provides protection from lock-in to proprietary vendor software. Using XML based orchestration business process execution language (BPEL) enables task sharing across multiple enterprises using a combination of Web services. Web services combine the advantages of the component-oriented methods and web techniques. Maintaining Web service quality requires more effort to manage overall Web service framework than each of Web service. Web Services Manager is a security administrator's environment designed to secure access to Web services and monitor activities performed on protected Web services. Web services provide platform independence for the service oriented communication. This way data integration can be done providing the service as a request and service as provider.

Keywords: Service Oriented Architecture (SOA), Web Services, Business Process Execution Language, XML.

1 Introduction

SOA is an architectural style whose goal is to achieve loose coupling among assorted interacting software applications. It also enables organizations to take advantage of existing investments in applications and systems. Using a SOA an organization can be more focused on resources and budget on innovation and on delivering new business services as re. SOA reuses services to automate a business process [1]. Services include Web Service Description Language (WSDL) for service interface definition and XML Schema Documents for message structure definition [3]. Instance of service is Web Services involves operation like Find, Bind, and Execute. Web Service is independent, modulated application which is described, published, located and called through the Internet. The architecture of SOA based on Web Service consists of three roles: service provider, service requester and service register [5].

Component Role in SOA

- **Service Provider:** The service provider creates a web service and publishes its interface and access information to the service registry. Service provider must decide which services to expose, how to make trade-offs between security of various services.

- **Service consumer:** The service consumer or web service client locates entries in the broker registry using various find operations and then binds to the service provider in order to invoke one of its web services.
- **Service Registry:** It registers what services are available with service provider, and lists all the potential service recipients. Before providing the service to consumer from provider, registry validates available services in the service registry.

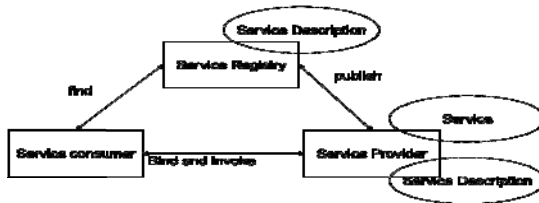


Fig. 1. SOA basic architecture specifying component

Advantages of using SOA

- Reduction in development time and cost
- Lower maintenance cost
- High-quality services
- Lower integration costs
- Reduce risk

Steps to deploy SOA [5]

- **Service enablement:** each discrete application needs to be exposed as a service.
- **Service orchestration:** Distributed services need to be configured and orchestrated in a unified and clearly defined distributed process.
- **Deployment:** Emphasis should be shifted from test to the production environment, addressing security, reliability, and scalability concerns.
- **Management:** Services must be audited, maintained and reconfigured. The latter requirements require that corresponding changes in processes must be made without rewriting the services or underlying application.

Web Services [3]

Web services using XML provide information to applications through an application oriented interface, so that it can be parsed and processed easily rather than being formatted for display. It publishes details of their functions and interfaces, keeping their implementation details private; thus a client and a Service interact regardless of the platforms on which they run or the programming languages in which they are written. Thus Web services applicable to a distributed heterogeneous environment.

A Web service is a software system designed to support interoperable machine-to-machine interaction over a network. [2]

1. The key specifications used by Web services

- eXtensible Markup Language - for formatting, exchanging structured data used in web services.

- SOAP (Simple Object Access Protocol)—an XML based protocol specifying envelope information, contents and processing information for a message.
- WSDL (Web Services Description Language)—an XML-based language used to describe the attributes, interfaces and other properties of a Web service. A WSDL document can be read by a potential client to learn about the service.

2. Agents and Services: A Web service must be implemented by a concrete agent. The agent is the concrete piece of software or hardware that performs the operation of sending and receiving messages. Here the service is the resource characterized by the abstract set of functionality provided by web service.

3. Requesters and Providers: A Requester entity’s Web service will use a requester agent to exchange messages with the provider entity’s provider agent. This is for the exchange of the messages.

4. Service Description: A Web service description (WSD) is a machine processable specification of the Web service’s interface, written in WSDL. It defines the message formats, data types, transport protocols, and transport serialization formats that should be used between the requester agent and the provider agent with the specification of one or more network locations at which a provider agent can be invoked, with information about the message exchange pattern that is expected.

5. Semantics: The semantics of a Web service is the shared expected result of the service, particularly in response to messages that are sent to it. This is the "contract" between the requester entity and the provider entity regarding the purpose and consequences of the interaction for communication or message exchange. The semantics represents a contract governing the meaning and purpose of interaction.

Overview of Engaging a Web Service [2]

- 1) The requester and provider entities become known to each other with their structure and working style.
- 2) The requester and provider entities agree upon standard protocol for communication, with the required service description and semantics.
- 3) The service description and semantics are realized by the requester and provider to their respective agents.
- 4) The requester and provider agents exchange messages as per the service requirement, performing some task on behalf of the requester and provider entities.

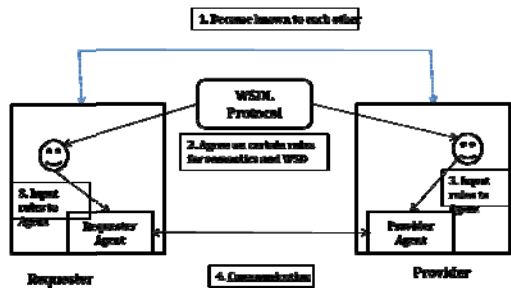


Fig. 2. Working of Web Services

SOA and Web Service [2]

SOA is a distributed systems architecture characterized by web service properties as:

- 1. Logical view:** The service is an abstracted, logical view of actual programs databases, business processes, etc. It is defined in terms of what it *does*.
- 2. Message orientation:** The Web Service defined as the messages exchanged between provider agents and requester agents, but not defines as the properties of the agents themselves. By avoiding any knowledge of the internal structure of an agent, one can incorporate any software component or application that can be "wrapped" in message handling code that allows it to adhere to the formal service definition.
- 3. Description orientation:** A service is described by machine-processable metadata which supports the public nature of the SOA. Details that are exposed to the public and important for the use of the service should be included in the description. The semantics of a service should be documented by its description.
- 4. Granularity:** Services tend to use a small number of operations with relatively large and complex messages.
- 5. Network orientation:** Services tend to be oriented toward use over a network, even though this is not an absolute requirement.
- 6. Platform neutral:** Messages are sent in a platform neutral, standardized format delivered through the interfaces. XML is the most obvious format that meets this constraint.

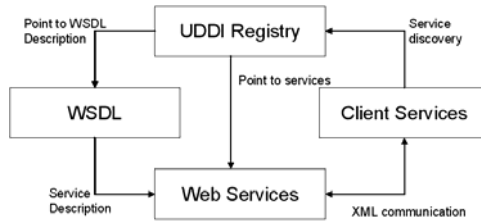


Fig. 3. Key techniques for web service

Operation mode of web services [6]

- Publish the service: Services provider publishes services WSDL description information in the UDDI register centre.
- Compose, or orchestrate, the services into business flows: Then through querying UDDI registering centre, services requester gains the services WSDL documents to provide the web services interoperability information.
- Services requester sends SOAP request to the services provider, and then services provider returns SOAP response messages to services requester.

Web services properties [11]

1. Discoverable
2. Communicable
3. Conversational
4. Secure and Manageable

Business Process Execution Language[8]

BPEL is an XML-based language for enabling task sharing across multiple enterprises using a combination of Web services. BPEL is based on simple objects access protocol and Web service description language. BPEL provides enterprises with an industry standard for business process orchestration and execution. Using BPEL, it is possible to design a business process that integrates a series of discrete services into an end-to-end process flow. This integration reduces process cost and complexity. The BPEL language enables to define how to: Send XML messages to, and asynchronously receive XML messages from, remote services. Manipulate XML data structures. Manage events and exceptions. Design parallel flows of process execution.

Features of BPEL [6]

- Rich sequencing semantics including parallel and asynchronous processing, Uses a compensation based long running transaction (LRT) model
- Provides rich scoped fault handling capabilities, Provides rich scoped asynchronous event handling capabilities allowing time based alerts as well as out of band events such as order cancellation
- Uses Web Services as the model for process decomposition and assembly; that is each BPEL process is a Web Service and can be composed of BPEL processes
- Uses XML and XPath for data access and manipulation.

Structure of a BPEL Process [8]

A business process described with <process> tag defines XML namespaces used during the description with a set of attributes.

```
<process name = NOName" targetNamespace="anyURL"
  queryLanguage="anyURL"? expressionLanguage="anyURL"?
  xmlns=http://docs.casispen.org/wsbpel/2.0/process/exetable>
  <import namespace="anyURL"?
  Location=anyURL'?
  importType="anyURL" />
<partnerlinks>...</partnerlinks>
<variable>...</variable>
<correlationSets>...</correlationSets>
Activity
</process>
```

BPEL Process Manager (with Human Workflow)

BPEL Process Manager provides a framework for easily designing, deploying, monitoring, and administering processes based on BPEL standards. It uses web service standards such as XML, SOAP, and WSDL. It handles Dehydration (enables the states of long-running processes to be automatically maintained in a database) and correlation of asynchronous messages. It performs parallel processing of tasks. It handles fault handling and exception management during both design time and run time. It checks event timeouts and notifications. It manages scalability and reliability of processes.

Overview of Workflow Services

- Workflow services enable to interleave human interaction with connectivity to system and services within an end-to-end process flow
- High level view of Workflow services in BPEL Process
- Task – Work needs to be done by user, role or Group
- Notification – An email, voice, fax, SMS that is send when user is assigned as task or informed that the status of the task has changed
- Work list – an enumeration of task or work item assigned to or of interest to a user
- Human Task Editor – A tool that enables to specify task setting such as task outcome, payload structure, notification setting and so on.
- Task file – The metadata task configuration file stores task setting specified with the Human Task Editor
- Routing slip – contains information about flow pattern for the Workflow, assignee, escalation policy, expiration duration, sequence in which participants interact in task.

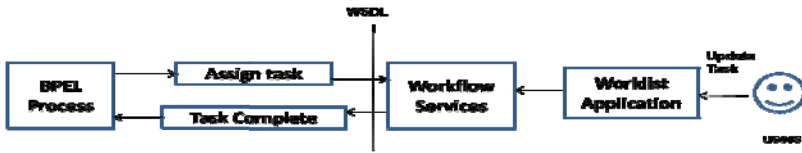


Fig. 4. Overview of workflow services

Conclusion

- Service-Oriented Architecture is an architectural style whose goal is to achieve loose coupling among diverse interacting software applications, enabling reusing services to automate a business process.
- System constructed on SOA using web services standards provide greater interoperability and some protection from lock-in to proprietary vendor s/w.
- Web services combine the advantages of the component-oriented methods and web techniques, and can describe their own services.
- SOA with Web Service based standards provide greater interoperability and some protection from lock-in to proprietary vendor software.
- BPEL is an XML-based language for enabling task sharing across multiple enterprises using a combination of Web services; BPEL provides enterprises with an industry standard for business process orchestration and execution. BPEL Process Manager provides a framework for easily designing, deploying, monitoring, and administering processes based on BPEL standards.

References

1. Radhakrishnan, R., Sriraman, B.: Aligning Architectural Approaches towards an SOA-Based Enterprise Architecture. In: Proc. Working IEEE/IFIP Conference on Software Architecture 2007 (WICSA), Mumbai, India, pp. 38–38 (January 2007)
2. Web Services Architecture W3C Working Group Note (February 11, 2004)

3. Srinivasan, I., Treadwell, J.: An Overview of SOA computing
4. Deng, W., Yang, X., Zhao, H., Lei, D., Li, H.: Study on EAI Based on Web Services and SOA. In: International Symposium on Electronic Commerce and Security
5. Erl, T.: Service-Oriented Architecture Concepts, Technology, and Design
6. Deng, W., Yang, X., Zhao, H., Lei, D., Li, H.: Study on EAI Based on Web Services and SOA. In: International Symposium on Electronic Commerce and Security 2008 (2008)
7. Securing web services & service oriented architectures with oracle web service manager 11g
8. Fabra, J.: BPEL2DENEb: Translation of BPEL Processes to Executable High-level Petri Nets. In: 2010 5th International Conference on Internet and Web Applications and Services (2010)
9. UDDI TC, Universal Description, Discovery and Integration v 3.0.2 (UDDI), <http://www.oasisopen.org/committees/uddispec/doc/spec/v3/uddiv3.0.220041019.htm>
10. Ortega, D., Uzcátegui, E., Guevara, M.M.: Enterprise Architecture and Web Services. In: 2009 Fourth International Conference on Internet and Web Applications and Services (2009)

A Meta Search Approach to Find Similarity between Web Pages Using Different Similarity Measures

Jaskirat Singh and Mukesh Kumar

University Institute of Engineering and Technology,
Punjab University, Chandigarh
jaskiratj@gmail.com
mukesh_rai9@yahoo.com

Abstract. Search engines are the online services available, which are used to locate necessary information on World Wide Web. As the web is growing at a very rapid rate, the pages that are similar to each other are also increasing. Hence, it is better to have a system that can discover similar web pages. In this paper, A Meta search approach is applied for the information retrieval purpose which retrieves pages from the result list of different search engines and content present in the web pages is analyzed on the basis of which system finds similarity between them. Web pages are represented in vector space which represents each web document as a vector and the terms present in that webpage as its components. Similarity is computed by using different similarity measures i.e. Cosine Similarity, Jaccards Coefficient and Dice Coefficient. A comparative analysis of these similarity measures is done to find out which measure performs better in terms of precision as well as recall.

Keywords: Meta Search, Vector Space Model, Similarity Measures, Tf-idf, Cosine Similarity, Jaccard Coefficient, Dice Coefficient, Link Structure.

1 Introduction

The World Wide Web has made a remarkable growth in its size due to rapid creation of loads of web pages containing both valuable (authoritative) pages and pages that refers or has links to these authoritative pages. This could be assured by the present size of the web which has grown to approx 60 billion pages. In order to retrieve information vested in these web pages, an information retrieval system is required. One such system is called a *Search Engine*.

The search engine allows the users to retrieve information from www by formulating the request in the form of a query. The search engines return the list of web pages in a manner that the web page with the high rating is listed at the top of the result. Now there are many measures opted for calculating the rating of web pages. One such popular measure is the calculation of *Page Rank* as suggested by Lawrence Page and Sergey Brin [1]. The web crawler and indexer are the main components of search engine. Web crawler collects the web pages from the web and indexer sorts these pages which are then used by the search engine. Search engines use query processors that process user queries and return matching answers in an order as determined by the ranking algorithm.

Now traditional web search engines take a query as input and produce a set of relevant pages. Relevant pages are the one that provide information which matches the interest of the user. This work proceed towards finding of 'similar' pages in the world wide web which involves finding other pages that address the same topic as the original page and also the information that they provide is also relevant. Search engines have a disadvantage that users have to formulate queries that satisfy their information needs, which is prone to errors. Hereby we apply a different approach to query a search engine i.e. instead of querying as a set of query terms; we input the search engine with address of a particular page i.e. URL of a particular page and it returns as output a set of related web pages. For example; if we input www.monster.com as query then the search engine should return other websites containing web pages related to jobs and recruitment on the web. This approach requires that the user has with him the page of interest. Also in this work the Meta Search Engine approach is applied for information retrieval. These are also called *Hybrid Search Engines*. In this the user runs single search and it searches different search engine databases on the web and then gives the best results on a single page. As each search engine has its own search criteria and its own database. Hence, meta search engines efficiently uses the resources of different search engine and produces a combined result. The results that appeared from this approach are then processed to determine which web pages are more similar to the queried page. Similarity is computed using different similarity measures using vector space model.

2 Problem Statement

As the web is growing at a rapid rate, hence the documents/pages that are similar to each other are also increasing. Presenting the same content on different webpage is purposely done so that the information is highly available and can be easily accessed by the users. This reduces traffic on busy websites. But giving the same content on different web pages has certain disadvantages like pages that are similar to each other are expensive in terms of crawling, computation time, storage and indexing. Duplicate pages affects the ranking of the true page as the authority scores (in-degree) gets divided among other pages. Web pages may contain content copied from other popular webpage i.e. high degree of plagiarism prevails in these pages. Big clusters are usually formed by pages that are similar to each other. Hence a system is needed that can efficiently determine similarity between web pages. Also different search engines (for example google: related) employ different techniques to retrieve web pages. Since their methodology being propriety, it is not known how they retrieve similar web pages. Usually the known approaches analyze either the link structure that exist between them or the content that is present between them.

3 Related Work

The different ways of finding similarity between web pages is achieved by analyzing the

1. *Content of web page*:- textual content, anchor text, meta tags (title, description, keywords).
2. *Link Structure of web page*:- distribution of internal and external links to the page under analysis.
3. *Content and Link both*:- considering both the information contained in content and how web pages are linked to each other.

To determine the structural similarity between the web pages, the *Tag Frequency Distribution Analysis(TFDA)* [16] measure is used which uses the frequency of HTML tags to calculate similarities. Different HTML tags holds the textual content of the webpage. **A. Tombros and Z. Ali [16]** gave an approach in which Tags are placed in classes depending upon their semantic connotations. Content within each individual class is indexed separately and treated as the representatives of each class. Indexed terms are then assigned different weights.

Vector Space Model [2] [17] In Vector Space model a set of documents is represented as vectors in a common vector space is known as the vector space model. The VSM assumes that for each term there exists a vector and then the linear combination of these term vectors represents documents in the vector space. Vector space model can be used to find the similarity between two documents by using various similarity measures such as cosine similarity, jaccards coefficient and dice coefficient. The vector space model converts text-based information to numerical vectors that are then used for analysis. A collection of n documents can be represented in vector space model by a term-document matrix. An entity in the matrix corresponds to the weight of the term in the document. *Tf-idf Weighting*:- used to produce the composite score for each document is described as.

$$Tf - idf = Tf_{t,d} \times idf_t, \quad Score(q, d) = \sum_{t \in q} Tf - idf_{t,d}$$

Once the vectors for all the documents are achieved, the similarity is calculated using a similarity measure. A similarity measure is a function that computes degree of similarity between two vectors. **Cosine Similarity[2] [12] [17]** is a way of calculating the similarity between two documents d_1 and d_2 i.e. by computing the cosine similarity of their vector representations $V(d_1)$ and $V(d_2)$

$$Sim(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)| |\vec{V}(d_2)|}$$

where the numerator gives the dot product of vectors, where as the denominator is the product of their Euclidean lengths. $V(d)$ denotes the document vector for d , with K components $V_1(d), \dots, V_K(d)$. The Euclidean length of d is calculated as $\sqrt{\sum_{i=1}^M \vec{V}_i^2(d)}$. When the value of cosine is 1, the vectors point in the same direction and when the value of cosine is 0, they are pointing in perpendicular directions, indicating dissimilarity between the documents. **Jaccard Coefficient [12][17] and Dice Coefficient** are also similarity measures that work on the vector space model. They differ from Cosine similarity in a manner that how they normalize their vectors. The jaccard coefficient and Dice coefficient between two vectors d_1 and d_2 is defined as

$$Jaccard\ Sim(d_1, d_2) = \frac{(d_1) \cdot (d_2)}{(|(d_1)| + |(d_2)| - (d_1) \cdot (d_2))}, \quad Dice(d_1, d_2) = \frac{2 \times N_{Common}}{N_1 + N_2}$$

Duplicates or Near-Duplicates[2] in web are the multiple copies of the same content. Some webpages are exact duplicates of each other due to mirroring and plagiarism. The search engines try to avoid indexing multiple copies of the same content in order

to keep down storage and processing overheads. The simplest approach to detect duplicates is to compute for each web page, a fingerprint that is a (k-bit) digest of the characters on that page. When the fingerprints of the two web pages are equal, then one of them is declared to be a duplicate of the other. Near-duplicates are the pages that differ by a close margin, say few characters. **Charikar’s SimHash [14]** is practical implementation for finding near-duplicates. *SimHash* is a fingerprinting technique that uses the property that fingerprints of near-duplicates that differ in small number of bit positions. To determine that Hamming distance between the two fingerprints is computed. **Gurmeet Singh Manku[15]** In their experiment validated that for a repository of 8 billion web pages, 64-bit simhash fingerprints are reasonable. *Fingerprints of Shingles* (which are the sequences of terms in a document d) of the document can also be computed to find the near duplicates. **Cocitation Algorithm [4]** The co-citation algorithm has its roots from finding similarity between the scientific papers. Two papers A and B are co-cited if a third paper C has citations to both of them. Co-citation can be applied to web documents by treating links as citations i.e. two web pages are co-cited if a third web page has links to both of them. Hence let K be a web page and D be the set of pages that link to K, called the parent of K. Then we say pages K1 and pages K2 are co-cited if more no of D1 and D2 are in common. Hence we can express similarity between pages K1 and K2 as:

$$Similarity(K1, K2) = \frac{D1 \cap D2}{D1 \cup D2}$$

The number of common parents of two nodes is called the degree of co-citation. The **Companion Algorithm [4]** is one of the approach which helps in finding related pages. This algorithm works on the principle of parent-child relationship. If there is a hyperlink from page P1 to P2 we say that P1 is a parent of P2 and page P2 is a child of page P1. The algorithm starts with an input query which is a URL of a page and finds similarity as shown in figure

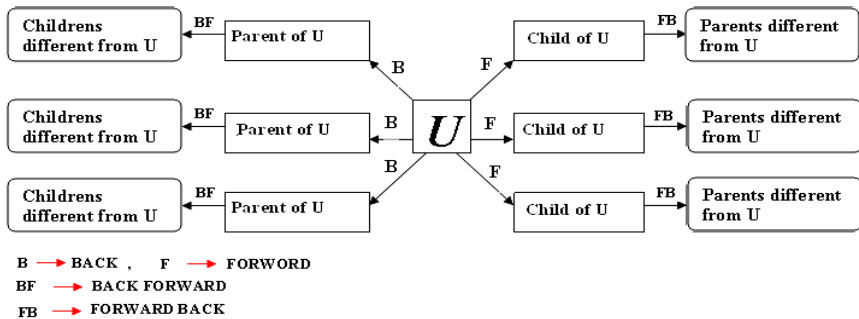


Fig. 1. The companion algorithm starts with an initial query page U and tracks back to find parents of U and forward to find childrens of U

Kessler[11] introduced the measure known as *bibliographic coupling* for determining similarity between scientific papers. Two documents are said to have a bibliographic coupling if they both cite a same paper. The basic idea is that the authors who work on the same subject tend to cite the same papers. This principle is

applied to web pages and states that the authors of the same subject tend to cite to the same pages.

$$\text{Bibliographic Coupling } (p1, p2) = \frac{(Cp1 \cap Cp2)}{|(Cp1 \cup Cp2)|}$$

Thus two pages have one unit of bibliographic coupling between them if they link to the same page. The more children in common page p1 has with page p2, the more related they are. **Amsler [10]** introduced a measure of similarity that combines both co-citation algorithm and bibliographic coupling. According to the approach two web pages P1 and P2 are related if

1. Same page links both webpage P1 and P2.
2. P1 and P2 link the same page.
3. P1 links a third page P3 which links the page P2

Hence more the links (either parents or children) P1 and P2 have in common more they are related to each other.

4 Proposed Work

In the Proposed work, meta search approach is applied for the retrieval of webpages from the World Wide Web. Since each search engine has its own propriety method by which they extract relevant web pages. So the power of these Search Engines to find web pages can be utilized to gather relevant webpages. For similarity evaluation and representation of webpages, Vector Space Approach is applied and then the similarity scores are computed for the intial web page and the pages extracted. First the initial collection is created that contains the Urls of webpages for which other similar web pages are to be found. One Url is selected from the initial collection and the keywords of that page are extracted and the Tf-IDF values are computed. These values obtained are the weights associated with the keywords and represents the composite score for the document. The keywords obtained are fetched to Google and Yahoo in Comma Separated Format and the respective search engines produces a result list according to the input. Each result list is stored separately and links present in them are extracted. The links that are present in both the result list are then selected once and a merged list is prepared that contains links from both the result lists. For each url in the merged list the keywords are extracted and again for those keywords the TF-IDf values are calculated. For the values obtained, Cosine Similarity, Jaccard Coefficient and Dice Coefficient scores can be calculated and analysed.

4.1 Algorithm

Step 1: Pick URL i from Initial Collection of URLs [1-n].

Step 2: Extract keywords present in that URL and calculate Tf-idf values.

```
Foreach(keyword in keywords[])
    Calculate Tf-idf(keyword)
```

Step 3: Send keywords in CSV format to Google and Yahoo as query.

- Step 4:** Extract the links produced in the result set of each search engine
 Result_Set1 \leftarrow Google;
 Result_Set2 \leftarrow Yahoo;
- Step5:** Merge the result sets produced in step 4 and eliminate Duplicate links
 Final_Result_Set \leftarrow Result_Set1 + Result_Set2;
- Step6:** Foreach(Link j in Final_Result_Set)
 Perform Step 2 again where Url = j
- Step7:** Compute Cosine Similarity
 Store Results
- Step8:** Compute Jaccards Coefficient
 Store Results
- Step9:** Compute Dice coefficient.
 Store Results
- Step10:** Comparative Analysis

5 Result and Discussion

In this work only the first ten results shown by each search engine i.e. Google and extracted. So in this case the maximum URL links that a merged list can have is 20. The table1 shows the values obtained through different similarity measures when the input query page D is *www.seo.com* arranged in an increasing order of their score values.

Query1: *www.seo.com*

Table 1. Similarity Scores for Query1 shows the values obtained through different similarity measures when evaluating the similarity for each page with the input query page D and are arranged in an increasing order of their score values

DOCUMENT	COSINE SIMILARITY	JACCARDS COEFFICIENT	DICE COEFFICIENT	AVERAGE SCORE
Sim(D,D9)	0	0	0	0
Sim(D,D2)	0.4437	0.1290	0.2285	0.2670
Sim(D,D3)	0.7016	0.1355	0.2388	0.3586
Sim(D,D8)	0.7393	0.1551	0.2686	0.3876
Sim(D,D5)	0.9413	0.3148	0.4788	0.5783
Sim(D,D13)	0.9413	0.4705	0.6400	0.6839
Sim(D,D16)	0.9626	0.4705	0.6400	0.6910
Sim(D,D12)	0.9726	0.4800	0.6486	0.7004
Sim(D,D7)	0.9786	0.5000	0.6666	0.7150
Sim(D,D6)	0.9805	0.6400	0.7804	0.8003
Sim(D,D15)	0.9922	0.8000	0.8888	0.8936
Sim(D,D11)	0.9922	0.8039	0.8913	0.8958
Sim(D,D4)	0.9922	0.9333	0.9655	0.9636
Sim(D,D10)	0.9938	0.9733	0.9864	0.9845
Sim(D,D1)	0.9998	0.9827	0.9913	0.9912

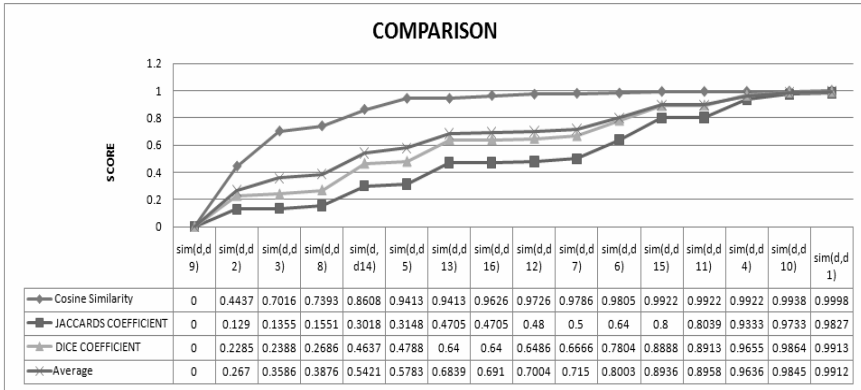


Fig. 2. Shows the comparison of similarity measures for query1 and the no of documents identified by cosine similarity are more as compared to other similarity measures

Query2 : www.espstar.com/other-sports/news/detail/item437780/India-beat-South-Korea-3-2-/

Table 2. Similarity Scores for Query2 shows the values obtained through different similarity measures when evaluating the similarity for each page with the input query page D and are arranged in an increasing order of their score values

Document	Dice Coefficient	Cosine Similarity	Jaccard coefficient	Average Score
Sim(D,D14)	0	0	0	0
Sim(D,D17)	0.0063	0.0799	0.0032	0.0298
Sim(D,D18)	0.0063	0.0799	0.0032	0.0298
Sim(D,D6)	0.1173	0.1288	0.0623	0.1028
Sim(D,D11)	0.1104	0.1928	0.0584	0.1205
Sim(D,D15)	0.2058	0.2101	0.1147	0.1768
Sim(D,D9)	0.1558	0.2253	0.0845	0.1552
Sim(D,D13)	0.1952	0.229	0.1081	0.1774
Sim(D,D12)	0.2286	0.2334	0.129	0.197
Sim(D,D8)	0.2214	0.254	0.1244	0.1999
Sim(D,D4)	0.03184	0.2826	0.0161	0.1101
Sim(D,D16)	0.3334	0.3787	0.2001	0.304
Sim(D,D10)	0.2594	0.4592	0.149	0.2892
Sim(D,D5)	0.3744	0.6485	0.2303	0.4177
Sim(D,D7)	0.1019	0.9043	0.0536	0.3532
Sim(D,D2)	0.939	0.9474	0.8851	0.9238
Sim(D,D3)	0.9722	0.9757	0.946	0.9646
Sim(D,D1)	0.9833	0.9841	0.9673	0.9782

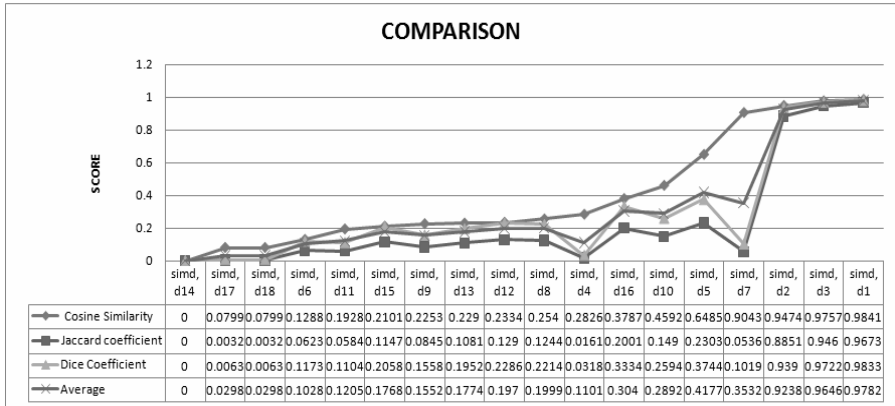


Fig. 3. Comparison of Similarity Measures for Query2

5.1 Validation of Results

The effectiveness of each similarity measure can be evaluated by calculation of two parameters namely Precision and Recall which are defined as:-

1. **Precision :-** Precision is defined as

$$\text{Precision} = \frac{\text{Number of similar webpages}}{\text{Total number of webpages}}$$

The threshold value of the angle between the vectors is kept at $\theta = 30^\circ$.

Table 3. Precision calculated for different similarity measures

QUERY/MEASURE	COSINE	JACCARD	DICE
Precision Query1	4/18 = 0.22	3/18 = 0.16	3/18 = 0.16
Precision Query2	9/16 = 0.56	2/16 = 0.12	3/16 = 0.18

The Precision is calculated for each similarity measure and as it can be analyzed from the figure 4 the Precision of Cosine Similarity is higher than that of Jaccard and Dice Coefficient for both the queries.

2 **Recall :-** Recall is defined as

$$\text{Recall} = \frac{\text{Number of similar webpages retrieved}}{\text{Total number of similar webpages}}$$

In this work, the total similar web pages present are determined by checking each web page whether they relate to the same topic and are of same type or not.

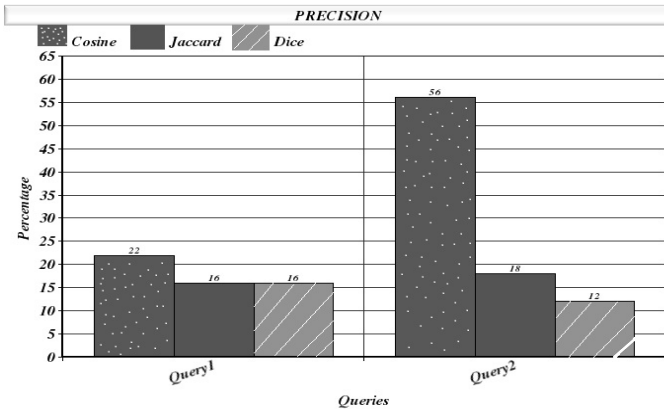


Fig. 4. Shows precision for cosine similarity is highest among all

Table 4. Recall Calculated for different similarity measures

QUERY/MEASURE	COSINE	JACCARD	DICE
Recall Query1	$8/11 = 0.72$	$2/11 = 0.18$	$3/11 = 0.27$
Recall Query2	$4/5 = 0.8$	$3/5 = 0.6$	$3/5 = 0.6$

The recall is calculated for each similarity measure and as can be analyzed from the figure 5 the recall of Cosine Similarity is higher than the Jaccard and Dice Coefficient for both the queries(query1 and query2). Hence, it is a better to use Cosine Similarity for determination of similar web pages as both the precision and recall for cosine similarity are higher than other similarity measures. Moreover the Meta Search approach can be incorporated with similarity calculation as it is an efficient approach for the retrieval of web pages from the web.

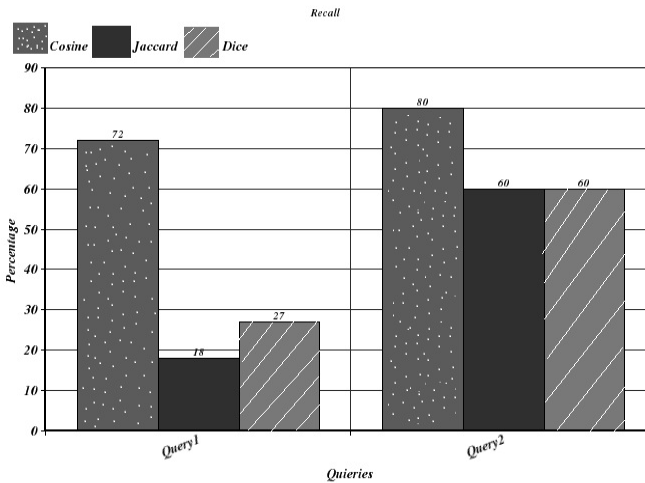


Fig. 5. Shows Recall of each similarity measure for different queries

References

1. Brin, S., Page, L.: The Anatomy of a Large Scale Hypertextual Web Search Engine. In: Proceedings of the Seventh International World Wide Web Conference, Brisbane, Australia (April 1998)
2. Manning, C.D., Raghavan, P.: An introduction to Information Retrieval. Preliminary draft© 2008 Cambridge UP (2008)
3. Kleinberg, J.M.: Authoritative Sources in a Hyperlink Environment. *Journal of the ACM, (JACM)* (1999)
4. Dean, J., Henzinger, M.R.: Finding Related Pages in the World Wide Web. In: The Proceedings of the 8th International World Wide Web Conference (May 1999)
5. Chirita, P.A., Olmedilla, D., Nejdil, W.: Finding Related Hubs and Authorities. In: The Proceedings of First Latin American Web Congress (2003)
6. Smucker, M.D., Allan, J.: Find-Similar: Similarity Browsing as a search tool. In: SIGIR 2006, pp. 461–468. ACM Press, New York (August 2006)
7. Grangier, D., Bengio, S.: Inferring Document Similarity From Hyperlinks. In: The Proceeding of CIKM 2005, pp. 359–360. ACM, New York (November 2005)
8. Fogaras, D.: Scaling Link based Similarity Search. In: The Proceeding of 14th International World Wide Web Conference, Japan (2005)
9. Lempel, R., Moran, S.: The Stochastic Approach for link-structure analysis (SALSA) and the TKC effect. In: The Proceedings of the 9th International World Wide Web Conference, Amsterdam, Netherlands (2000)
10. Amsler, R.: Application of citation-based automatic classification. Technical report, the University of Texas at Austin Linguistics Research Center (December 1972)
11. Kessler, M.M.: Bibliographic Coupling Between Scientific Papers. *American Documentation* (1963)
12. Srikant, R., Bayardo, R.J., Ma, Y.: Scaling Up All Pairs Similarity Search. In: The Proceedings of 16th International Conference on World Wide Web, Canada (May 2007)
13. Di Iorio, E.: Detecting near –replicas on the Web by content and hyperlink analysis. In: The Proceedings of International Conference on Web Intelligence (WI 2003). IEEE, Los Alamitos (2003)
14. Charikar: Similarity Estimation Techniques from Rounding Algorithm. In: The Proceedings of the 34th Annual ACM Symposium on Theory of Computing. ACM Press, New York (2002)
15. Manku, G.S., Jain, A., Sarma, A.D.: Detecting Near- Duplicates for Web Crawling. In: The Proceedings of International Conference on World Wide Conference, Canada (May 2007)
16. Ali, Z., Tombros, A.: Factors affecting Web Page Similarity. In: Losada, D.E., Fernández-Luna, J.M. (eds.) ECIR 2005. LNCS, vol. 3408, pp. 487–501. Springer, Heidelberg (2005)
17. Nagwani, N.K., Bhansali, A.: An Object Oriented Email Clustering Model Using Weighted Similarities between Emails Attributes. *The Proceedings of International Journal of Research and Reviews in Computer Science, IJRRCS* (January 2010)
18. Aslam, J.A., Frost, M.: An Information–theoretic Measure for Document Similarity. In: The Proceedings of 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Canada (July 2003)
19. Shivakumar, N., Garcia-Molina, H.: Finding replicated web collections. In: The Proceedings of International Conference on Management of Databases, on Research and Development in Information Retrieval, Canada (July 2003)

20. Salton, G.: A Vector Space Model for automatic Indexing. *Communications of ACM* (November 1975)
21. Tsatsaronis, G., Panagiotopoulou, V.: A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness. In: *The Proceedings of EACL Student Research Workshop, Athens*, pp. 70–78 (April 2009)
22. Simmetrics, Open Source API for text similarity Measurement,
<http://www.dcs.shef.ac.uk/~sam/simmetrics.html>

Cost Estimation for Distributed Systems Using Synthesized Use Case Point Model

Subhasis Dash and Arup Abhinna Acharya

School of Computer Engineering
KIIT University, Bhubaneswar, Orissa, India
subhasisbbsr@gmail.com,
arupacharya.kiit@gmail.com

Abstract. Cost Estimation for distributed systems is a major challenge nowadays. Estimating the cost of development for distributed systems is based on a prediction of the size for future systems. A lot of cost estimation models were reported in the literature but many of these models became obsolete because of the rapid changes in technology. Reliable estimations are difficult to obtain because of the lack of detailed information about the future system at an early stage as well as due to the distributed location of various components of the developed software. Cost models like COCOMO(COnstructive COst MOdel)[5,6] and sizing methods like Function Point analysis are well known and in widespread use in Software Engineering. These models were applicable only to procedural paradigm, and are not directly applicable to software products developed using the object oriented methodology or distributed systems. It is this idea that gave birth to the creation of Use Case Point (UCP) metrics, originally developed by Gustav Karner[3]. UCP uses use cases as the primary factor, use case model is the first model developed in an object-oriented design process using UML. In this paper we extend the UCP to estimate the cost of development for distributed systems. We propose a novel approach to map the distributed systems from their function points and converting use case point counts on the basis of actor interaction with the actors present at other locations to the software and to estimate cost of development by using Distributed Synthesized UCP (ds-UCP) model with additional information obtained from distributed synthesized use case attributes.

Keywords: Use Case Point, COCOMO, Cost Estimation, Distributed Systems.

1 Introduction

Software is an intangible product and hence probably the most crucial difference between the manufacturing industry and the software industry is that the former is able to stick to schedules and cost most of the time. Even when using a well-defined methodology, the development cost of a well-defined application is not easy to predict. Some key factors that contribute to this difficulty include the precise set of functionalities to be implemented, the various risks associated with the development process, the knowledge

and experience of the development team and lack of detailed information about the system at an early stage. Among these, the set of functionalities to be implemented is the most crucial factor. Software cost estimation models developed in the 60's and 70's used the Line of Code (LOC) or Delivered Source Instruction (DSI) metrics. For example, COCOMO model [5, 6] used the DSI metrics while many others used LOC metrics. The major problem with these metrics was two-fold: (i) the lack of precise definition of LOC or DSI, and (ii) there is no reasonable methodology by which one can estimate the number of source code lines or instructions until the coding is over.

Gustav Karner [3] came up with the notion of Use Case Point (UCP) which is somewhat similar to the notion of Function Point but based on use cases. Model based estimation will help in a greater extend to reverse engineering and maintenance of legacy software. The use case model is the front end model of the Unified Modeling Language (UML) [9]. However, Karner's [3] method does not take into account some of the application domain details such as the number of interaction between actors, and relationships between use cases.

The objective of this research is to elaborate the UCP model by synthesizing the use cases with a focus on internal details of each use case as well as interaction between the use cases present at various locations. It is common that in a distributed system every use case is supported by a use case distributed synthesized attributes that explains the internal details of the use case. This paper describes a method for cost estimation based on the use case diagram at an early stage of software development. It focuses on the use case attributes, uses the interaction between the entities in a use case diagram, and hence closely estimates the size of the software product to be developed. Both direct and indirect use case dependencies between the use cases present at various locations are taken into account to give a more accurate result. The methodology was tested with the use cases developed for an automated indexing and searching system for a book publishing company. The estimated value for development efforts seems to match with the actual development efforts spent by the company.

The rest of the paper is organized as follows: Section 2 describes the step by step application of the synthesized UCP methodology. A case study with a step wise evaluation of UCP using ds-UCP model is described in section 3. The paper concludes in Section 4 with the discussion on continuing work in this direction in Section 5. Due to space constraints, this paper does not include descriptions of a use case model such as notations and their semantics. Interested readers are referred to any UML book such as [10] or UML manual published by OMG [9].

2 Proposed Method: Distributed Synthesized UCP (ds-UCP)

The ds-UCP methodology uses every aspect of a use case model such as actors, use cases, interaction between actors and use cases present at distributed locations, relationships between actors, relationships between use cases and finally the distributed synthesized attributes of each use case. The last one is important because it describes the missing details of a use case diagram, while others can be directly extracted from a use case diagram itself. This makes the significant difference between UCP method given by Karner [3] and ds-UCP. The attributed list also contains few use case narratives suggest by Periyasamy [4]. In order to concretize the methodology, the authors used the template for use case synthesized attributes in Table 1.

Table 1. Sample Use Case Distributed Synthesized Attributes

Use case name	A descriptive name of the use case
Purpose	A brief description of the tasks to be implemented by this use case
Input parameters	List of Input Parameters to the use case
Output parameter	List of Output parameters returned from the use case.
Primary actor	The list of actors who invoke this use case
Secondary actor	The list of actors that are used by this use case
Precondition	Conditions that must test true to use the use case. Unlike assumptions, these conditions are tested by this use case before doing anything else. If the conditions are not true, the actor or other use case is refused entry.
Post condition	Conditions that must test true when the use case ends. You may never know what comes after the use case ends, so you must guarantee that the system is in a stable state when it does end.
Process	A step-by-step description of the dialog between the use case (the system) and the user (actor or other use case). Very often it is helpful to model this sequence of events using a flowchart or activity diagram just as you might model a protocol for communication between two business units.
Successful scenario	A sequence of instruction that explain the successful scenario of invoking this use case.
Exceptions	A set of conditions that may make the use case fail when invoked.
Includes List	Use cases that can be included in this use case.
Extends List	Use cases that can be extended from this use case.
Dependency-I	Dependency between Use cases present at individual site.
Dependency-II	Direct dependency between Use cases present at different sites
Dependency-III	Indirect dependency between Use cases present at different sites.

2.1 Use Case Estimation

The ds-UCP method first calculates unadjusted UCP values, referred to as UUCP in this paper. This includes unadjusted actor weights, unadjusted use case weights. The unadjusted UCP values are then adjusted using three additional factors, namely technical complexity factor (TCF), environment factor (EF), and distributed synthesized factor (DSF). The final result is the adjusted UCP values, referred to as AUCP. The ds-UCP method uses the same TCF and EF weights as given in the original UCP method [3], but it uses a different set of calculations for UCP values. The authors of this paper believe that the final UCP value (ds-UCP) returned by AUCP is more precise compared to those returned by UCP.

2.1.1 Actor Weight Classification

Table 2 shows the assignment of weights to various actors based on the actor type and the number of transactions performed by an actor.

Table 2. Actor Weight Classification

Actor type	Classification of actors	Weight
Very simple	Specialized Primary/Secondary actor	1.0
Simple	Simple Primary actor with $1 < \text{number of transactions} \leq 3$	1.5
Less average	Primary actor with $3 < \text{number of transactions} \leq 5$	2.0
Average	Primary actor with number of transactions > 5	2.5
	Secondary actor with 1 transaction	2.5
Complex	Complex Secondary actor with $1 < \text{number of transactions} \leq 3$	3.0

This, in turn, is based on the number of transactions the actor has with the use cases. As the number of transactions increases, more effort is involved in coding and hence the complexity of the corresponding actor increases. The complexity of a secondary actor (such as a database) with ‘n’ transactions is higher than that of a primary actor (such as a user) with the same number of transactions because generally more effort is involved in coding transactions with the secondary actor compared to those

with a primary actor. For example, if the secondary actor is a database, then coding efforts are required for checking the connection to the database, writing SQL statements for transactions, checking for database commit actions and so on.

2.1.2 Use Case Weight Classification

Similar to actors, each use case is assigned a different weight. The use case type is defined based on the number of transactions (the number of direct connections between actors and the use case). Table 3 lists the classification of use cases and their associated weights [7].

Table 3. Use Case Weight Classification

Use case type	Classification of use cases	Weight
Simple	Number of transactions <= 3	5
Average	4 < number of transactions <= 7	10
Complex	7 < number of transactions <=	15

2.1.3 Weights for Synthesized Use Case Attributes

A use case diagram must be supported by Synthesized Use case Attributes. Table 1 shows the structure of a synthesized use case attributes used in this methodology. Though there is no standard for the structure of a use case attributes, the authors found that the use case structure illustrated in Table 1 contains all information that many practitioners use. With this assumption, Table 4 describes the weights associated with the different parameters of a use case attributes.

Notice that all the use case attributes shown in Table 1 are used in Table 4 because all of them contribute to the coding efforts, and others such as actors are already taken into consideration. Moreover, it should also be noted that the weight associated with ‘Precondition’ in Table 4 must be used for only one predicate in the precondition; the same applies to post condition as well.

Thus, a complex precondition such as

Bank account number must be valid ^ PIN code must be valid

will include two simple predicates. The justification of assigning separate weight for each individual predicate comes from the fact that each simple predicate is required to be implemented to validate the precondition.

Table 4. Weights for Use case Distributed Synthesized Attributes

Attribute Number	Use case Synthesized Attributes	Weight
DS1	Input parameter	0.1
DS2	Output parameter	0.1
DS3	A condition to execute each process	0.1
DS4	A predicate in Precondition	0.1
DS5	A predicate in Post-condition	0.1
DS6	An action in Successful scenario	0.2
DS7	An exception	0.1
DS8	An Includes conditions	0.1
DS9	An Extends conditions	0.1
DS10	Dependency between Use cases present at individual site.	0.1
DS11	Direct dependency between Use cases present at different sites	0.2
DS12	Indirect dependency between Use cases present at different sites.	0.3

3 Evaluation of ds-UCP: A Case Study

In the original method the total unadjusted actor weight (UAW) is calculated by counting how many actors there are of each kind (by degree of complexity), multiplying each total by its weighting factor & adding up the products. Each use case is then defined as simple, average complex depending on number of transactions in the use case descriptions, including secondary scenarios.

3.1 Proposed Estimation Method

The primary goal of proposed technique is to estimate the cost of software from number of different actors and its use case interactions both at individual site and at distributed site. In this proposed technique, all the actors are classified into very simple, simple, less average, average, complex and then the use case are classified into simple, average and complex based on the number of actor interacting with that use case along with use case interacting with another use case. Hence use case complexity can be defined easily. The ds-UCP method also uses distributed synthesized weight factor (DSF) to improve the effectiveness of this technique.

Further classification is made for all the use cases that interact with another use case. If the current use case is interacting with a simple use case then we will classify it as average and if the current use case is interacting with an average use case then we will classify it as complex. Then we will assign the weighting factor in the above said manner.

3.1.1 Estimating ds-Ucp

Finally, the synthesized use case point (ds-UCP) is calculated by multiplying all the three values UUCP, TCF, EF & DSF. That is,

$$ds-UCP = UUCP * TCF * EF * DSF \tag{1}$$

CASE STUDY: - A Book Indexing Project

Given below is the use case diagram of indexing project for which we will calculate the use case points.

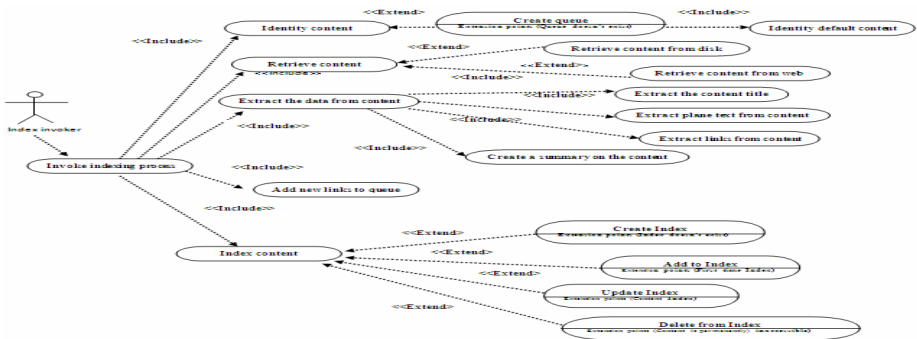


Fig. 1. Use case diagram for book indexing project

3.2 Weight Based Estimation

The method employs a technical factors multiplier corresponding to the technical complexity adjustment factor of the FPA method, environmental factors multiplier in order to quantify non-functional requirement & synthesized factor in order to quantify the use cases. Various factors influencing productivity are associated with weight and values are assigned to each factor, depending on the degree of influence.

The Technical Factor (TCF) is calculated multiplying the value of each factor (T1 – T13) by its weight and then adding all these numbers to get the sum called the TFactor. Finally, the following formula is applied:

$$TCF = 0.6 + (0.01 * TFactor) \tag{2}$$

The Environmental Factor (EF) is calculated accordingly by multiplying the value of each factor (F1 – F8) by its weight and adding all the products to get the sum called the EFactor. The formula below is applied:

$$EF = 1.4 + (-0.03 * EFactor) \tag{3}$$

The Distributed Synthesized Factor (DSF) is calculated accordingly by multiplying the value of each factor (S1 – S9) by its weight and adding all the products to get the sum called the SFactor. The formula below is applied:

$$DSF = 1.1 + (0.01 * SFactor) \tag{4}$$

3.2.1 Weight Based on Technical Factor

The technical factors and their weighted values are shown in the following table:-

The TCF valued is calculated using equation (2):

$$TCF = 0.6 + (0.01 * 40) = 1$$

Table 5. Technical Factors in project and their Weights

Factor Number	Description	Weight	Assessment	Impact
T1	Distributed System	2	0	0
T2	Response adjectives	1	4	4
T3	Response adjectives	1	5	5
T4	Complex processing	1	2	2
T5	Reusable code	1	3	3
T6	Easy to install	0.5	4	2
T7	Easy to use	0.5	4	2
T8	Portable	2	4	8
T9	Easy to change	1	3	3
T10	Concurrent	1	3	3
T11	Security features	1	3	3
T12	Access for third parties	1	1	1
T13	Special training required	1	4	4
T-FACTOR		TOTAL WEIGHT		40

3.2.2 Weight Based on Environmental Factor

The environmental factors and their weighted values are shown in the following table:-

Table 6. Environmental Factors in project and their Weights

Factor Number	Description	Weight	Assessment	Impact
F1	Familiar with RUP	1.5	2	3
F2	Application experience	0.5	1	0.5
F3	Object-Oriented experience	1	2	2
F4	Lead analyst capability	0.5	1	0.5
F5	Motivation	1	2	2
F6	Stable requirements	2	3	6
F7	Part-time workers	-1	0	0
F8	Difficult programming language	-1	4	-4
E-FACTOR		TOTAL WEIGHT		10

The EF value is calculated using equation (3):

$$EF = 1.4 + (-0.03 * 10) = 1.1$$

3.2.3 Weight Based on Synthesized Factor

The DSF value is calculated using equation (4):

$$DSF = 1.1 + (0.01 * 2.0) = 1.12$$

The UUCP value is calculated using the following equation:

$$UUCP = UAW + UUCW \tag{5}$$

$$UUCP = 1*2 + 15*5 + 3*10 + 0*15 = 107$$

Using the above values the use case points of this project is calculated using equation (1):

$$ds-UCP = 107 * 1 * 1.1 * 1.12 = 131.824$$

Table 7. Weights for Use case Distributed Synthesized Attributes

Attribute Number	Use case Synthesized Attributes	Weight	Assessment	Impact
DS1	Input parameter	0.1	2	0.2
DS2	Output parameter	0.1	2	0.2
DS3	A condition to execute each process	0.1	1	0.1
DS4	A predicate in Precondition	0.1	1	0.1
DS5	A predicate in Post-condition	0.1	1	0.1
DS6	An action in Successful scenario	0.2	1	0.2
DS7	An exception	0.1	1	0.1
DS8	An Includes conditions	0.1	1	0.1
DS9	An Extends conditions	0.1	1	0.1
DS10	Dependency between Use cases present at individual site.	0.1	1	0.1
DS11	Direct dependency between Use cases present at different sites	0.2	2	0.4
DS12	Indirect dependency between Use cases present at different sites.	0.3	1	0.3
S-FACTOR		TOTAL WEIGHT		2.0

3.3 Result Analysis

The UCP count using ds-UCP method will give more accurate and effective estimation as compared to the UCP counts done by Bente Anda [2] and Periyasamy [4]. The comparison made is listed in Table 8.

Table 8. Comparative Study

SL. NO.	METHOD	UCP Count
1	Bente Anda [2]	107.23
2	Periyasamy [4] (e-UCP)	118.91
3	ds-UCP(Proposed method)	131.824

4 Conclusion

This paper evaluates a use case diagram for an indexed project and calculated the synthesized use case points (ds-UCP) according to proposed method which help us to calculate the cost of the project. Here the proposed method constructed a simple way to use minimal set of formula to calculate the cost software development for distributed systems. This motivates a relationship between some of the existing use case methods. With classifying the actors here, the approach was able to classify the software system with respect to the level of code quality and the better way to calculate the use case points.

This approach is very promising and more applicable to identifying low quality code than high and use case points due to specific theoretical concept employed to develop the approach. However, this is a mammoth step in the right direction in reducing the turn-around time. It takes to perform a code analysis on industrial software cost estimation.

5 Continuing Work

The ds-UCP method currently uses the technical complexity factors and environmental factors as defined by Karner [3]. The authors have planned to revise these factors in order to improve the calculations. Development of concurrent and real time systems is a challenging job now a day. So, the author also planned to extend this ds-UCP method for estimation of cost for development of concurrent and real time systems. The ds-UCP method may further be extended to prioritize the distributed synthesis attributes by which we can get a better picture of estimation at an early stage.

References

- [1] Albrecht, A.J.: Measuring application development productivity. In: IBM Applications Development Symp., GUIDE Int. and SHARE Inc., IBM Corp., Monterey, CA, October 14-17 (1979)
- [2] Anda, B., et al.: Estimating software development efforts based on use cases - Experience from industry. In: Gogolla, M., Kobryn, C. (eds.) UML 2001. LNCS, vol. 2185, p. 487. Springer, Heidelberg (2001)
- [3] Karner, G.: Metrics for Objectory. Diploma thesis, University of Linköping, Sweden. No. LiTHIDA- Ex- 9344:21 (December 1993)
- [4] Periyaswamy, K., Ghode, A.: Effort Cost Estimation using extended Use Case Point (e-UCP) Model. IEEE, Los Alamitos (2009)
- [5] Boehm, B.: Software Engineering Economics. Prentice Hall, Englewood Cliffs (1981)

- [6] Boehm, B., et al.: Software Cost Estimation with COCOMO II. Prentice Hall, Englewood Cliffs (2000)
- [7] Edward, C.R.: Estimating Software Based on Use Case Points. In: Proceedings of the Object-Oriented, Programming, Systems, Languages, and Applications (OOPSLA) Conference, San Diego, CA (2005)
- [8] Fetke, T., Abran, A., Nguyen, T.: Mapping the OO-Jacobsen approach into function point analysis. In: Proceedings of Technology of Object-Oriented Languages and Systems, (1997)
- [9] UML 2.0 Reference Manual, Object Management Group (2003),
<http://www.omg.org>
- [10] Schneider, G., Winters, J.P.: Applying Use Cases, 2nd edn. Addison Wesley, Reading (2001)

Comparative Approach to Cloud Security Models

Temkar Rohini

Department of MCA,
Vivekananda Institute of Technology, Mumbai 77, India
rohini_dighe@rediffmail.com

Abstract. With today's increasing trend of cloud computing world, most of the organizations prefer to outsource and share their data by the means of cloud service providers. Cloud computing plays a major role by providing different resources in the form of web services which is based on pay-as-per-usage model. Along with benefits of reduced cost, dynamic resource availability, consumption based cost it also brings new challenges for data security and access control when users outsource sensitive data for sharing on cloud servers. This paper takes a glance on the different cloud computing model based on their access and deployment of clouds, also it provides two cryptographic approaches for the data security. The first approach is software based data security model and another is hardware based data security model. In software based model the cryptographic approach is used to build the software for the security of data. In hardware based model it is embedded in the hardware itself, thus providing more robustness. Though hardware based model provides more robustness; the software based model provides more flexibility for the data correction and data recovery.

Keywords: Cloud Computing, Cloud Security Model, Cryptography.

1 Introduction

Cloud computing is Internet-based computing, whereby shared resources, software, and information are provided to computers and other devices on demand, like the electricity grid. The basic cloud computing architecture is illustrated in fig 1. Usually cloud computing services are delivered by a third party provider who owns the infrastructure. Moving data into the cloud offers great convenience to users since they don't have to care about the complexities of direct hardware management. As a result, cloud computing gives organizations the opportunity to increase their service delivery efficiencies, streamline IT management and better align IT services with dynamic business requirements. In many ways, cloud computing offers the "best of both worlds," providing solid support for core business functions along with the capacity to develop new and innovative services. From the perspective of goodness cloud computing has following characteristics:

- Pay as you go – payment is variable based on the actual consumption.
- Highly abstracted – server hardware and related network infrastructure is highly abstracted from the users.

- Virtual – Physical location and underlying infrastructure are transparent to users.
- Multi-tenant – multi-tenant architectures allow numerous customer enterprises to subscribe to the cloud computing capabilities while retaining privacy and security over their information.
- Immediately scalable – usage, capacity, and therefore cost, can be scaled up or down with no additional contract or penalties.

Google Apps, Google App Engine, Amazon Web Services are some examples of trustworthy cloud service providers.

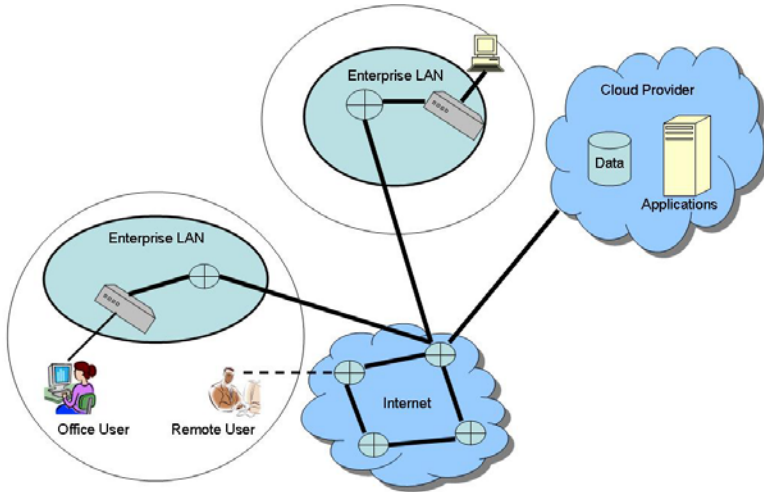


Fig. 1. Basic Cloud Computing Architecture

2 Cloud Architectural Models

Basically cloud computing has two types architectural models. First is based on service provided by cloud is Service Model. Second is based on access by client is Deployment Model.

2.1 Cloud Service Models

1. Infrastructure as a service (IaaS): where a customer makes use of a service provider's computing, storage or networking infrastructure.
2. Platform as a service (PaaS): where a customer leverages the provider's resources to run custom applications. PaaS enables you to create web applications quickly, without the cost and complexity of buying and managing the underlying software/hardware.
3. Software as a service (SaaS): where customers use Software that is run on the providers Infrastructure. Software as a Service is an application operation where a provider would issue a license to a client for the use of software.

2.2 Cloud Deployment Models

1. **Private Cloud:** In a private cloud, the infrastructure is managed and owned by the customer and located on-premise. In particular, this means that access to customer data is under its control and is only granted to parties it trusts.
2. **Public cloud:** In a public cloud the infrastructure is owned and managed by a cloud service provider and is located off-premise. This means that customer data is outside its control and could be granted to untrusted parties.
3. **Hybrid Cloud:** A hybrid cloud is a cloud computing environment in which an organization provides and manages some resources in-house and has others provided externally. Ideally, the hybrid approach allows a business to take advantage of the scalability and cost-effectiveness that a public cloud computing environment offers without exposing mission-critical applications and data to third-party vulnerabilities.

Apart from these basic cloud deployment Models we have one more deployment model called as community cloud where infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party and may exist on premise or off premise.

3 Cloud Threats

Apart from all the advantages of cloud service, cloud data security is the main issue of quality of service. Organizations have to deal with in order to isolate their data from other cloud clients and to fulfill confidentiality and integrity demands. Cloud Computing is not just a third party data warehouse. The data stored in the cloud may be frequently updated by the users, including insertion, deletion, modification, appending, reordering, etc. Thus along with secured data storage on demand dynamic data storage is equally important issue. The threats to information assets residing in the cloud can vary according to the cloud delivery models used by cloud user organisations. Here are some examples of cloud threats.

- 1) **Insider User Threats:** malicious cloud provider user, malicious cloud customer user, malicious third party user, malicious third party user are the examples of insider user threats.
- 2) **Data Leakage:** Failure of security access rights across multiple domains and failure of electronic and physical transport systems for cloud data and backups can lead to data leakage.
- 3) **Data Segregation:** Incorrectly defined security perimeters and incorrect configuration of virtual machines cause data segregation.
- 4) **User Access:** Implementation of poor access control procedures creates many threat opportunities, for example that disgruntled ex-employees of cloud provider organisations maintain remote access to administer customer cloud services, and can cause intentional damage to their data sources.

4 Software Based Cloud Security Models

The Overview of Software Based Cloud Security Model is illustrated in Fig 2. The security model is similar to the antivirus software, which automatically encrypts the file before storing on the cloud. Three different entities can be identified as follows:

- User: users, who have data to be stored in the cloud and rely on the cloud for data computation, consist of both individual consumers and organizations.
- Cloud Service Provider (CSP): a CSP, who has significant resources and expertise in building and managing distributed cloud storage servers, owns and operates live Cloud Computing systems.
- Third Party Auditor (TPA): an optional TPA, who has expertise and capabilities that users may not have, is trusted to assess and expose risk of cloud storage services on behalf of the users upon request.

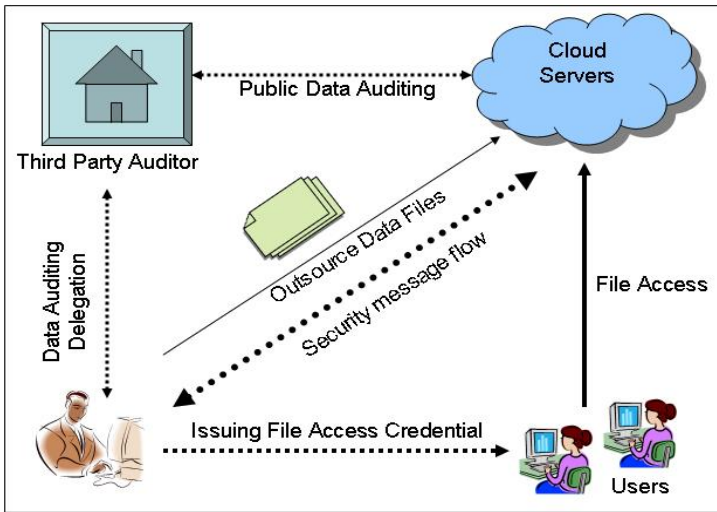


Fig. 2. Software Based Cloud Security Model

File Distribution Preparation:

- **F** – the data file to be stored. We assume that **F** can be denoted as a matrix of m equal-sized data vectors, each consisting of l blocks. Data blocks are all well represented as elements in Galois Field $GF(2^p)$ for $p = 8$ or 16 .
- **A** – The dispersal matrix used for Reed-Solomon coding.
- **G** – The encoded file matrix, which includes a set of $n = m + k$ vectors, each consisting of l blocks. The data file **F** redundantly across a set of $n = m + k$ distributed servers. A $(m + k, k)$ Reed-Solomon erasure-correcting code is used to create k redundancy parity vectors from m data vectors in such a way that the original m data

vectors can be reconstructed from any m out of the $m + k$ data and parity vectors. By placing each of the $m + k$ vectors on a different server, the original data file can survive the failure of any k of the $m+k$ servers without any data loss. The systematic layout with parity vectors is achieved with the information dispersal matrix \mathbf{A} , derived from an $m \times (m + k)$ Vandermonde matrix:

$$\begin{pmatrix} 1 & 1 & \dots & 1 & 1 & \dots & 1 \\ \beta_1 & \beta_2 & \dots & \beta_m & \beta_{m+1} & \dots & \beta_n \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \beta_1^{m-1} & \beta_2^{m-1} & \dots & \beta_m^{m-1} & \beta_{m+1}^{m-1} & \dots & \beta_n^{m-1} \end{pmatrix}$$

where $\beta_j (j \in \{1, \dots, n\})$ are distinct elements randomly picked from $GF(2^p)$. After a sequence of elementary row transformations, the desired matrix \mathbf{A} can be written as

$$\mathbf{A} = (\mathbf{I}|\mathbf{P}) = \begin{pmatrix} 1 & 0 & \dots & 0 & p_{11} & p_{12} & \dots & p_{1k} \\ 0 & 1 & \dots & 0 & p_{21} & p_{22} & \dots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & p_{m1} & p_{m2} & \dots & p_{mk} \end{pmatrix}$$

By multiplying \mathbf{F} by \mathbf{A} , the user obtains the encoded file: $\mathbf{G} = \mathbf{F} \cdot \mathbf{A}$.

To achieve assurance of data storage correctness verification tokens are computed. Before file distribution the owner pre-computes a certain number of short verification tokens on individual vector $G_{(j)} (j \in \{1, \dots, n\})$. Each token covers a random subset of data blocks. Whenever the user wants to make sure the storage correctness for the data, he challenges the cloud servers with a set of randomly generated block indices. Upon receiving challenge, each cloud server computes a short “signature” over the specified blocks and returns them to the user. The values of these signatures should match the corresponding tokens pre-computed by the user. Meanwhile, as all servers operate over the same subset of the indices, the requested response values for integrity check must also be a valid codeword determined by secret matrix \mathbf{P} [2].

5 Hardware Based Cloud Security Model

D-DOG (*Data Division and Out-of-order keystream Generation*), a high performance hardware implementation oriented stream cipher for distributed storage network. performance gap is through hardware implementation [3].

5.1 Divide and Store

Figure 3 illustrates the basic divide-and-store principles of the D-DOG scheme. At the local user side, the major functions include the following. When a data file is stored:

- 1) Generating the IV using three elements: the PIN from the user, the nonce generated by the system and the bits abstracted from the plaintext;
- 2) Constructing the keystream for encryption.
- 3) Encrypting the remained plaintext using the key stream.

- 4) Dividing the cipher text into multiple data blocks with fixed-length, the last block will be stuffed if it consists of fewer bits.
- 5) Allocating storage nodes and sending each block to one of them.
- 6) Storing the PIN & Keystream, and publish nonce.

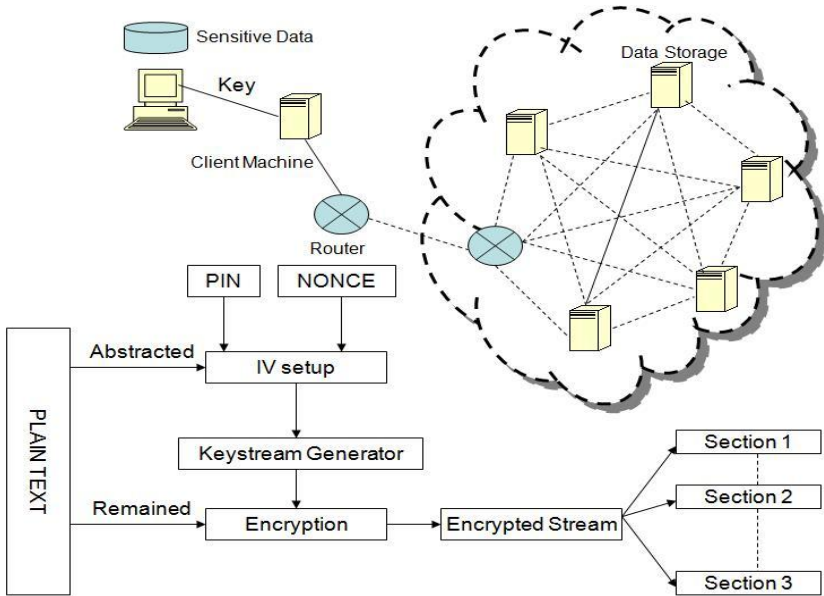


Fig. 3. D-DOG Scheme

5.2 Encryption and Decryption Using D-DOG

This scheme encrypts the plaintext and decrypts the ciphertext by performing the bitwise adding calculation with a keystream. Figure 4 and Figure 5 present the flowcharts of the encryption operation and decryption operation respectively. The plaintext is inputted into the Separation module, which takes the pseudo-random stream generated by RandomAddrGen1 as the address index and draws the corresponding bits from the plaintext. KeystreamGen makes use of the Key and the data from IV Initial as input and output keystream. The Separation module outputs two separate streams: stream1 and stream2. Then, they are encrypted by the exclusive or with the keystream. After the stream1 and stream2 are encrypted, they can be combined together, or they can be sent out directly. As marked with a shadowed background, both the Combining module and the RandomAddrGen2 are optional. If the combining module is used, the RandomAddrGen2 is used to produce the address for the bit insertion operation. For module Keystream Generator2, there are three encryption pseudo-random generator schemes according to the Figure 4, which are corresponding to different decryption methods, as shown in Figure 5.

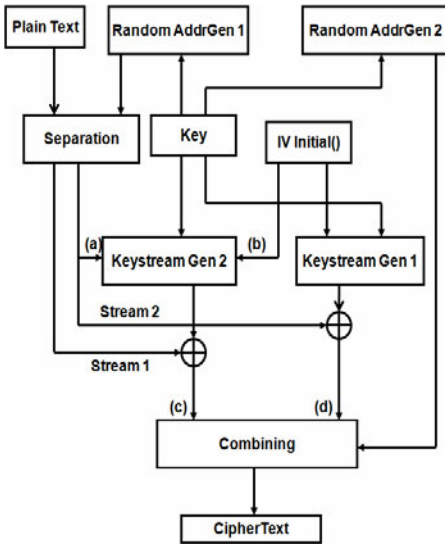


Fig. 4. Encryption Flow

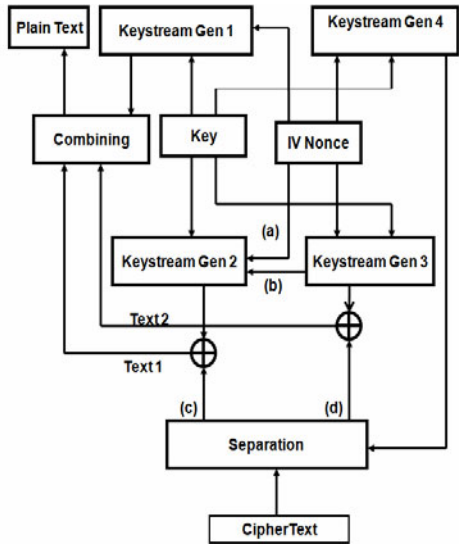


Fig. 5. Decryption Flow

6 Comparisons of Cloud Security Models

The following Table 1 show the comparative study of Software based and hardware based cloud security models.

Table 1. Comparisons of Cloud Security Models

	Software Based Cloud Security Model	Hardware Based Cloud Security Model
Network Topology(WAN)	Security system is not enough robust.	Security system robust.
Performance	Performance is lesser as compared to hardware based model.	Performance is higher.
Third party Auditor	Optional	Not Required
Recovery	Along with error detection it can also easily implement algorithm for correction and recovery	For recovery more hardware is required.
Cost	Implementation cost is less.	Implementation cost is more.

7 Amazon Simple Storage Service (Amazon S3) Security

With any shared storage system, the biggest question is whether unauthorized users can access information either intentionally or by mistake. Amazon S3 also uses the software based approach where the file is encrypted before storing on the cloud. With any shared storage system, the biggest question is whether unauthorized users can access information either intentionally or by mistake. Customers may wish to secure data even when it is being stored within Amazon S3. Data stored within Amazon S3 is not encrypted at rest by Amazon Web Services (AWS). However, users can encrypt their data before it is uploaded to Amazon S3 so that the data cannot be accessed or tampered with by unauthorized parties [6].

8 Conclusions

Cloud Computing is a vast topic. The above report gives a brief introduction to the cloud computing concept and its security issues. Cloud Computing brings the various characteristics like agility, reduced cost, device and location independence, security, metering. Cloud computing comes with two architectural models, one is service model and another is deployment model. Service models are Software as a Service, Platform as a Service, and Infrastructure as a Service. Deployment models are also of three types, private cloud, public cloud and hybrid cloud. Along with benefits of reduced cost, dynamic resource availability, consumption based cost it also brings new challenges for data security and access control when users outsource sensitive data for sharing on cloud servers.

To deal with the ever growing threat of cloud data we have two types of cloud security models. Software based cloud security models compute tokens for data verification. Hardware based cloud security Model uses key stream based cryptography to ensure storage security. Though hardware based model provides more robustness and performance, software based model provides more flexibility for data correction and recovery.

References

1. Kamara, S., Lauter, K.: Cryptographic Cloud Storage. Microsoft Research.
2. Wang, C., Wang, Q., Ren, K., Lou, W.: Ensuring Data Storage Security in Cloud Computing. IEEE, Los Alamitos (2009)
3. Feng, J., Chen, Y., Ku, W.-S., Su, Z.: D-DOG: Securing Sensitive Data in Distributed Storage Space by Data Division and Out-of-order keystream Generation. In: IEEE ICC 2010 Proceedings (2010)
4. IBM Perspective on Cloud Computing (2008)
5. Wang, C., Ren, K.: Toward Publicly Auditable Secure Cloud Data Storage Services. IEEE Network (2010)
6. http://s3.amazonaws.com/aws_blog/AWS_Security_Whitepaper_2008_09.pdf

Development of Agile Security Framework Using a Hybrid Technique for Requirements Elicitation

Sonia and Archana Singhal

Department of Computer Science, University of Delhi, Delhi, India
soniacsit@yahoo.com, singhal_archana@yahoo.com

Abstract. Today's competitive market demands immediate attention on security issues for developing secure software system. Security must be an integral part of any application development methodology. It becomes more challenging when developers design projects according to agile methodology. Traditional ways of development are sequential considering major changes during analysis. Agile methodology is required as there is a need for an iterative approach which encourages changes in requirements at any stage in software development life-cycle. In this paper, we are presenting a framework which effectively implements security practices in agile development and adopts additional features proposed by other researchers. The key point in our framework is that we are embedding a hybrid technique for requirement elicitation with Agile Software Development (ASD). This technique would combine abuser stories and attack trees drawing best features of each of their individual methods. This hybrid technique maps security threats found during security requirements effectively as compared to individual techniques.

Keywords: Agile Software Development, Software Security, Threat Modeling, Abuser stories, Attack Trees.

1 Introduction

During past decade, Agile methods such as Extreme Programming(XP), Scrum, Feature driven development, Adaptive software development have come into existence to overcome perceived and actual problems faced by prescriptive process models. But as explained in some research papers [5, 19] implementation of security is not as effective in them as needed. All proposed methods of agile methodology are based on some general principles defined by agile alliance [21] and the manifesto for ASD [20]. This includes iterative development, short development life cycles, emphasis on direct communication between customer and developer rather than heavy documentation, small focused teams and working software after each iteration.

For a qualitative system security must be treated as a highest priority to protect it against security threats. It can be achieved by mapping security practices into agile development. To build secure software systems using agile methods, it is necessary to incorporate security engineering into system design as soon as possible. It will eliminate need for complex problem to be secured in haphazard manner during later stages. As defined in [10], Systems Security Engineering is concerned with identifying security

risks, requirements and recovery strategy. For achieving this Threat Modeling act as foundation for specification of these security requirements. However threat modeling requires complete modeling and documentation efforts which are not favored by agile manifesto. But according to Microsoft SDLC/Agile, threat modeling process in ASD should be time boxed and limited to only parts of product that currently exist. It minimizes the developer time as required in ASD and provides benefits of threat modeling. This may not require security expertise.

In present paper we are integrating security practices where iterative development of software is not being overlooked. For this most important security requirements are handled first and the others in later releases. Our Agile Security Framework (ASF) presents some key features.

- It maps security requirements using a hybrid technique. This technique combines abuser stories with attack trees giving benefits of both techniques at a time for security requirement elicitation. Also limitations of these two can be fulfilled by each other.
- ASD releases projects in short iterations called sprints that are, from one week to two months. To implement all security requirements in single iteration is not possible. As according to [6] iterative architecture is one that develops with the system and includes only features that are necessary for current iteration or delivery. Therefore, we categorize the abuser stories resulting into development of a security framework. It helps us in knowing which story is to be considered in which iteration based on their risk impact, severity and user's need.

Although introducing security won't be an easy task during ASD but to some extent security engineering benefits from iterative development allowing gradual discovery of security issues and progressive implementation of its countermeasures.

Our paper is organized as follows. Section 2 gives the description of related work. Section 3 provides an overview of Agile Development Process and Security Requirements Engineering. Section 4 presents some key points for providing iterative development to our Agile Security Framework (ASF). Our proposed Agile Security Framework is given in Section 5 with conclusion and future work in Section 6.

2 Related Work

Many researchers have contributed in various ways to integrate security and agile processes [2, 13, 15, 18]. Danier Mellado has given a comparative study of proposals for establishing security requirements for development of secure software system [12]. Howard Chivers had delivered a quality work on agile security using incremental security architecture and also showed agile development for secure web applications by integrating risk assessment with agile processes [9, 14]. Similarly Beznosov has classified security assurance methods and techniques with regard to their clash with agile development [8]. Some papers have worked on use of attack trees associated with security requirement [16, 17]. John Peeters has put forward the idea of using abuser stories explaining how attacker may abuse the system and jeopardize stakeholder's assets with usual user stories ranked according to perceived threats [4, 12]. Vidar Kongsli has also given security in web applications using misuse

stories with user stories to capture malicious use of attacks [15]. They all suggested various techniques to implement security in agile processes. Need for further refinement in this area provoked us to develop Agile Security Framework (ASF). As multiple methods to specify security requirements exist but none of the above tries to create a complete framework for agile development in which they overlap abuser stories and attack trees to map security requirements. In our ASF we put together various techniques suggested by some researchers as required that provide step by step guidance to agile software developers. Focusing on iterative nature of agile development we also categorize security stories in different groups which helps us to know which security practice is to be implemented in which iteration.

3 Background

3.1 Agile Software Development (ASD)

Current era requires software development at fast pace and ever-changing. Building software with agile methodology is a great way to overcome this problem. Agile methods encourage customer satisfaction, iterative and incremental development, less documentation, informal nature and simplicity in development. Traditional design methods based on sequential lifecycle and having a predictive approach are often inconsistent with our modern needs. However in comparison to them ASD is adaptive in nature supporting continuous changes and produces working software in short duration.

ASD presents a number of methods and Extreme Programming (XP) is one of the most common methods of agile process. In this method functionality is prioritized by customer. It contains four activities consisting of Planning, Designing, Coding and Testing. We are broadly following phases of XP for our framework but certain changes are inevitable, due to security implementation. Although we are completely following agile principles to develop our proposed ASF described in Section 5.

3.2 Security Requirements Engineering

At time of creation software development should be done with security in mind at all stages and it should not be an afterthought. It includes measures to be taken throughout the software development lifecycle to prevent vulnerabilities of the system that creep in during requirement, design, development, deployment, upgrades or maintenance stages of system. Security engineering process comprises of various stages as explained below.

Threat Modeling. It identifies all possible threats after finding assets and access points of the system.

Risk Assessment. It assesses risks from threats identified by threat modeling and then prioritizes them for mitigation.

Requirements Elicitation. It can be achieved by various techniques like Misuse cases, Abuser stories, Attack trees, Softgoal interdependency graph. These different approaches have some benefits and limitations and every approach lacks proper

completeness. Therefore we need mechanisms for security requirements elicitation that will be palatable to regular software developer and suitable for use in ASD.

4 Key Points for Providing Iterative Development to Our ASF

As we are going towards the development of a framework for ASD, issues related to its short release cycle or iterations can't be neglected during security engineering. In each iteration the implementers must have space and reason to consider security [6]. In our proposed ASF, we suggest some points to be considered during software development.

- Here in first iteration we develop a security framework which categorizes abuser stories based on their severity, impact, risk factor and user's need. It is also discussed earlier by Baskerville et al. 2003 [14] that when the development is carried out in several development releases, the developers should be informed in which release the abuse case is prevented (i.e. countermeasure is implemented). By grouping them into different categories, developer can more effectively understand which story is to be implemented in which iteration. So instead of focusing on overall system he just plans for current iteration which helps him to achieve particular release on time securing well defined features. Keeping this in mind, developer applies threat modeling only on stories involved in current iteration that saves time and agility.
- In this framework, initial iteration starts with release planning that defines scope of whole project involving security requirements of complete system. This would be longer than that of successive iterations as understanding, planning and categorization of security features for further iterations will make it lengthier and time consuming. Instead of making first iteration longer developer can also devote complete second iteration for security implementation making all iterations equal in length. Successive iterations will just involve current iteration planning keeping watch on security stories assigned to them. In every iteration during the planning phase stories designed in previous iteration are reconsidered and updated if necessary according to current scenario. It might add new abuser stories.

5 Agile Security Framework (ASF)

In this Section we are presenting an iterative framework in which security is addressed at every stage of development lifecycle not considered an afterthought and agility is maintained by providing flexibility in implementing changes at any stage. Our ASF helps developers by providing step by step guidance in applying security techniques that will bring them to achieve a secure software system. The approach presented here gets its idea from mainstream agile development process and we mould it to inject some best security techniques in it. It includes hybrid technique (Combination of abuser stories and attack trees), Security Training Practice, Security Framework (Categorizing security requirements to be considered in different iterations).

Now we can enumerate ASF taking all aforementioned aspects into account. We propose this framework in defined phases although there boundaries are roughly

defined due to informal nature of ASD. These phases with complete structure are illustrated in detail in Fig. 1 and explained below.

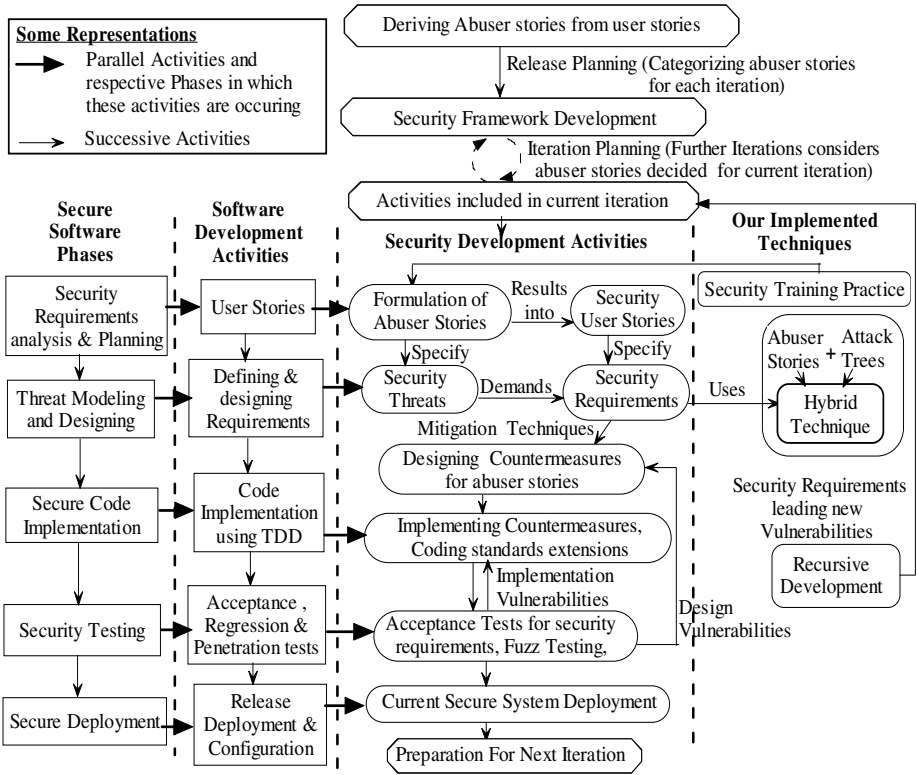


Fig. 1. Overall Structure of our Agile Security Framework (ASF)

5.1 Phase 1: Security Requirements Analysis and Planning

Critical assets identification. This phase explores user stories describing features and functional requirements of the software to be built. Keeping security in mind, developer will identify critical assets of the system for describing threats in future and in addition addresses specific security objectives like goals, constraints with user stories.

Formulation of abuser stories. When all assets and security objectives have been identified developer collects some of the abuser stories (describing undesired behavior of the system) from stakeholders on index cards. These abuser stories are based on given user stories and assets of the given system and can be seen as agile counterparts of abuse cases or misuse cases. After that remaining abuser stories must be created by developers and security experts based on their past experience. They analyze potential threats that could result in high risks to the assets using Threat Modeling process, which is used to shape up a secure software design. Threat modeling is a process by

which possible threats, attacks and vulnerabilities against the functionalities of software can be depicted. Developers also consider important security features required by customers for formulating abuser stories.

We emphasize here on Security Training Practice suggested by Xiaocheng Ge. et al. [1] for gathering these abuser stories. The given practice motivates developers and stakeholders to understand how to write abuser stories and helps in exploiting common security attacks, vulnerabilities, threats and risks.

5.2 Phase 2: Threat Modeling and Designing

Designing security requirements is a challenging activity and must be performed with great care and clarity. Overall Goals for security design process includes threat analysis, techniques to manage and mitigate risks and finally translate security requirements into reality (serves as a guide for implementation).

Risk Assessment and Prioritization. Threats identified in abuser stories possessing more risk will be implemented first. So after getting abuser stories they are prioritized based on their risk factor calculated by its impact on system. According to [7] simplest way to prioritize threats is by using two factors: damage and Likelihood (that means how much damage a risk can cause and likelihood of their occurrence). First calculate overall risk factor for each threat, and then sort the threat list by decreasing order of risk. Risk can be managed by risk acceptance, risk transfer, risk removal and risk mitigation. Security measures should also consider cost of recovery as it must not exceed cost of risk. Finally a selected threat for mitigation is applied in security requirements lifecycle. Sometimes, we will also develop some security user stories identified to mitigate some threats posed by abuser stories easily as shown in Fig. 1.

Requirements Elicitation using a Hybrid Technique. Implementing hybrid technique for security requirements elicitation in ASD is one of the key features of our framework. Here we are mapping selected threats into security requirements. Identification of potential threats and attacks along with vulnerabilities and its prioritization has been done in previous phase. Now with the help of hybrid technique suggested in [3] we will map these threats into security requirements in agile methodology. Our hybrid technique is more effective than other methods of security requirement elicitation including common criteria, misuse stories, attack trees and many more. The given approach as in [3] combines the strengths of misuse case and attack trees in traditional development methods. But as we are using this technique in ASD which encourages lightweight, ever-changing working software in small time constraints, so instead of misuse cases we propose here to combine abuser stories with attack trees.

Although both misuse cases and abuser stories describes how users can misuse a system with malicious intent, thereby used for identifying various security requirements. But major difference between these methods of requirement capture is the level of detail (or precision) in their respective textual descriptions [11]. Misuse Cases are heavier providing much information and require more time for analysis and writing thus completes in several iterations. Contrary to it, abuser stories of our ASF provides brief description of security requirements and informal in nature. These stories are comprehensive and encourage completion in single iteration. It doesn't model

interaction between actor and system and consider only those features of misuse cases that are required for our current iteration. In our approach we outline role, goal, precondition and mitigation conditions for our abuser stories leaving other features of misuse cases.

Mainly during mapping of threats we have to focus on all entry and exit points from misuser perspective (Points from which attacker try to harm the system) as mis-user always try to attack on system from points other than the entry points addressed by user. Therefore for complex abuser stories it will not be possible to get how the attacks in the given system can be realized which we can get by attack trees. But attack trees do not specify strategy for avoiding the threats and preconditions present at time of threat, which can be achieved by abuser stories. So, given hybrid technique provides benefits of both techniques simultaneously eliminating shortcomings of each other giving one-size-fits-all solution. For implementing this technique, our Hybrid diagram named as HDESIRE (Hybrid diagram embedding security illustrating requirements) is used to represent threats for security requirements elicitation. Various steps for creation of HDESIRE with overall process for security requirement elicitation using hybrid technique are illustrated in Fig. 2.

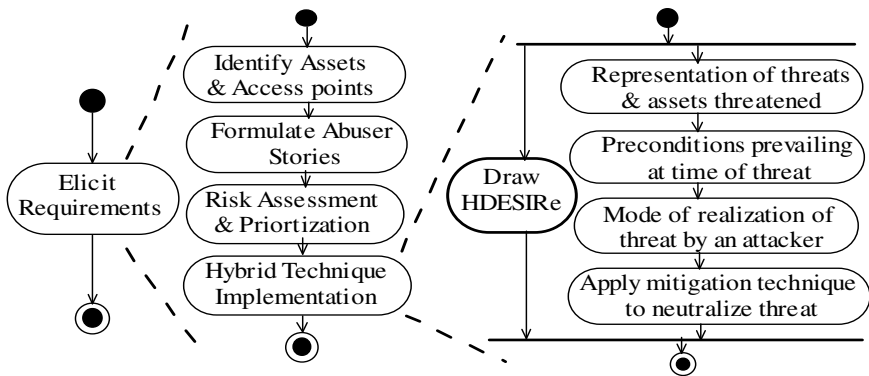


Fig. 2. Activity Diagram illustrating Security requirements elicitation method

Designing Security Requirements. By designing security requirements we mean that all security issues considered during planning phase should be modeled in it and its representation must be as simple as possible. As mentioned in Section 4 threat modeling must be implemented on security requirements and changes arise in current iteration. But sometimes designing these security requirements will lead to new vulnerabilities. For example, during an ATM attack abuser wants to obtain pin number of user to get unauthorized access of user’s account which can be mitigated by encrypting pin number during Iteration 1. But fulfilling the given security requirement by encrypting pin number will lead to new vulnerabilities like

- Encryption scheme applied in system can be guessed by malicious user.
- Encryption scheme or algorithm can be leaked by some insider.
- Encryption software gets destroyed and pin number entered, transmits online without encryption.

Given situation must be handled recursively in successive iterations as shown in Fig.1.

An Example illustrating HDESIRE for eliciting Security Requirements. Most of the network technologies, without integrating with security mechanisms originally, have to be redesigned to provide some security services. Automated Teller Machine (ATM) is one of those technologies. To build an ATM security system, according to our ASF first step is to find assets. Then we formulate abuser stories based on threats possessed by it and identify different attack paths from attacker's view. Then these threats are prevented by specifying security requirements for each threat.

Using a concrete example now we are sketching our hybrid technique that describes and documents security requirements. ATM will suffer a lot of threats like shoulder surfing, skimming, phishing, Denial of service, Man in middle attack and traffic analysis. From number of threats at different paths we will include some main threats. But these have sufficient complexity to explain our approach. Here our assets are money and user's secret information corresponding to his ATM card and pin number. The threat we are choosing for it is that "malicious user wants to steal money from the ATM". Formulated abuser story corresponding to given threat is

"An unauthorized user captures identification and authentication of authorized user for stealing money when authorized user taking out money from his account using ATM machine. It can be mitigated by protecting secret information".

Given abuser story is complex and require detailed step by step realization of threat which provides better understanding and clarity. It leads to new threats and hence more mitigation techniques are required to remove them. So we are drawing suggested HDESIRE to specify security requirements corresponding to each threat. Here root node represent attacker's goal present in our abuser story and attack tree as well, so act as a common point for combining the two. From given abuser story we can define preconditions under which given attack can be realized. Precondition is that "User is taking out money from his ATM account". It is shown on left side and assets on right side of root node. Leaf node of given root node in tree structure gives different ways of achieving the attacker's goal. Satisfying a tree node represents either satisfying all leaf nodes (AND) or satisfying a single leaf (OR). From the given diagram, mitigation conditions to remove the identified threats have been evolved by abuser stories. These are represented at the top of root node as shown in Fig. 3. One mitigation technique can mitigate several threats simultaneously so we can say mitigation techniques overlap each other. Due to this reason in the given example we have not shown individual mitigation technique for each threat.

Using HDESIRE given in Fig. 3 security requirements for the given problem is -

- Enforcing strong pin number policy will restrict unauthorized access of account and from brute force attack (Guesses Pin number).
- Privacy enhanced protocols will provide resistance from any type of information disclosure, thus provide protection from getting cloned or forged ATM card.
- Protecting secret data will guard against the misuse of pin number and card both. From this misuser can not get authentication by any means.
- By encrypting information misuser will not be able to capture ATM data from online banking site and further threats of its leaf nodes.

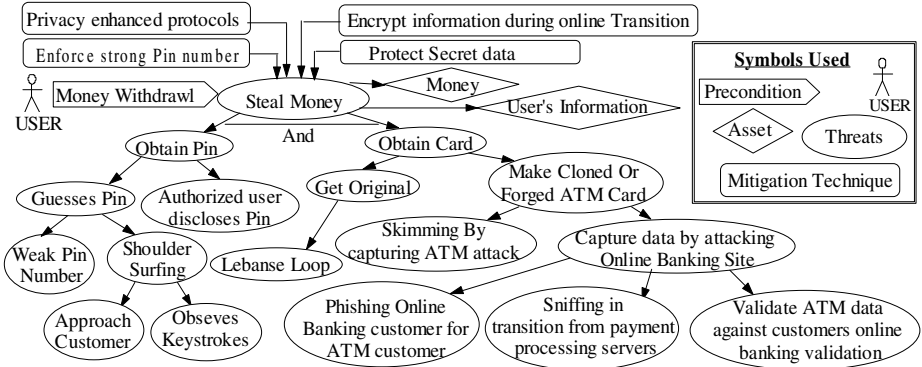


Fig. 3. Hybrid diagram embedding security illustrating requirements (HDESIRE) representing threat of stealing money from ATM

We have seen that how given HDESIRE is helpful in specifying security requirements. Now in agile process given security requirements will be expressed as acceptance tests. These acceptance tests ensure that a requirement under test works correctly. Proposed framework will help developers to figure out security requirements completely for designing and development of the given system in an informal agile environment.

5.3 Phase 3: Secure Code Implementation

During this phase implementation of security requirements and its countermeasures is performed in the same way as other requirements considered during development. In agile processes we will use Test driven development (TDD). In TDD series of unit test are included for the security stories of current sprint (iteration). It provides further security to developer as after that user will focus on what must be implemented to pass unit test. Also unit test developed in this phase, are not required to be generated again during testing. Here feedback from the project and actual realization of security analysis and designing comes in front of us. At the end of it this phase we will get the final implemented version.

5.4 Phase 4: Security Testing

Since we are considering security issues in ASF from the beginning thus they are aligned with our proposed agile framework at every stage. Testing should not wait till the end, it should be applied constantly and effectively to get a qualitative secure software systems. Various tests coming under the scope of secure system development are described below.

Unit Testing. Unit test creation before coding is a key element for testing. These unit test are implemented using automated testing so can be executed easily as many times as required and also helpful in regression testing. With that by automatic testing continuous integration has been possible which considers problems instantly wherever it appears.

Acceptance testing. Acceptance tests are derived from abuser stories in the same way as from user stories. The main goal here is to test the security requirements of abuser stories.

Above mentioned tests are applied in their respective phases. Testing applied in given phase includes

Fuzz testing. It is applied to find software problems by adding invalid unexpected data to an input using fuzz tester tool to see if the secure software fails.

Penetration Testing. This is performed to find vulnerabilities in some complicated applications.

5.5 Phase 5: Secure Deployment

In this phase software is ready for current release as each iteration results into working software. After deploying it current iteration is kept under observation to find out remaining vulnerabilities in the system. Preparation for next release also starts which includes computation of project velocity on the basis of number of stories implemented in previous release.

6 Conclusion and Future Work

With a history of low success rate using traditional development processes, it might not be wrong to say that an agile process is worth trying. There is a growing recognition of agile processes for system development. But achieving security satisfaction plays a critical role towards this. This paper modifies agile processes to address important security issues. Here our proposed ASF presents guidelines to develop adaptive, iterative software introducing security practices at each phase and we suggest security training for all developers and stakeholders. This paper also outlines a technique for eliciting security requirements overlapping abuser stories with attack trees resulting in an approach adopting best characteristics of these two methods and avoiding many of the toxins that make our approach anti-agile.

Security in agile development process becomes a widespread issue. A wide variety of methods were proposed presenting a general purpose of security requirements management but no individual method provides complete solution. Therefore there arises a need for a model that could contribute best features of already defined methods comprising a successful end result. We have made a modest attempt towards this. The future work includes a deeper study on the advanced risk analysis methods to derive a general structure for validating security requirements in agile framework.

References

1. Ge, X., Paige, R.F., Polack, F., Brooke, P.: Extreme Programming Security Practices. In: Concas, G., et al. (eds.) XP 2007. LNCS, vol. 4536, pp. 226–230. Springer, Heidelberg (2007)
2. Siponen, M., Baskerville, R., Kuivalainen, T.: Integrating security into agile development methods. In: 38th Annual Hawaii International Conference on System Sciences (2005)

3. Gandotra, V., Singhal, A., Bedi, P.: Identifying Security Requirements Hybrid Technique. In: Proceedings of the 4th International Conference on Software Engineering Advances, Porto, Portugal, pp. 407–412. IEEE Computer Society, Los Alamitos (September 2009)
4. Peeters, J.: Agile Security Requirements Engineering. In: Requirements Engineering for Information Security (2005)
5. Beznosov, K.: Extreme Security Engineering: On Employing XP Practices to Achieve 'Good Enough Security' without Defining It. In: First ACM Workshop on Business Driven Security Engineering (BizSec), Fairfax, VA (October 31, 2003)
6. Chivers, H., Paige, R.F., Ge, X.: Agile security using an incremental security architecture. In: Baumeister, H., Marchesi, M., Holcombe, M. (eds.) XP 2005. LNCS, vol. 3556, pp. 57–65. Springer, Heidelberg (2005)
7. Myagmar, S., Lee, A.J., Yurcik, W.: Threat Modeling as a Basis for Security Requirements National Centre for Supercomputing Applications. Univ. of Illinois at Urbana-Champaign (2005)
8. Beznosov, K., Kruchten, P.: Towards Agile Security Assurance. In: The New Security Paradigms Workshop, White Point Beach Resort, Nova Scotia, Canada, September 20-23 (2004)
9. Ge, X., Paige, R.F., Polack, F., Chivers, H., Brooke, P.J.: Agile Development of Secure Web Applications. In: ICWE 2006, July 11-14. ACM, New York (2006)
10. Kotonya, G., Sommerville, I.: Requirements Engineering: Processes & Techniques. John Wiley & Sons, Chichester (1998)
11. Davies, R.: The power of stories. WWW Retrieved, Citeseer (2001)
12. Mellado, D., Fernández-Medina, E., Piattini, M.: A Comparative Study of Proposals for Establishing Security Requirements for the Development of Secure Information Systems. In: Gavrilova, M., et al. (eds.) ICCSA 2006. LNCS, vol. 3982, pp. 1044–1053. Springer, Heidelberg (2006)
13. Daud, M.I.: Secure Software Development Model: A Guide for Secure Software Life Cycle. In: International MultiConference of Engineers & Computer Scientists, Hong Kong (2010)
14. Baskerville, R., Levine, L., Pries-Heje, J., Ramesh, B., Slaughter, S.: Is Internet speed Software Development Different? IEEE Software 20(6), 102–107 (2003)
15. Kongsli, V.: Towards Agile Security in Web Applications. In: The Proceedings of OOP-SLA, Portland, Oregon, USA, October 22-26. ACM, New York (2006)
16. Schneier, B.: Attack trees: Modeling Security Threats. Dr. Dobbs' Journal
17. Mauwl, S., Oostdijk, M.: Foundations of Attack Trees. P.1/32 (September 26, 2005)
18. Boström, G., Wäyrynen, J., Bodén, M., Beznosov, K.: Extending XP Practices to Support Security Requirements Engineering. In: SESS 2006, Shanghai, China, May 20-21 (2006)
19. Wäyrynen, J., Bodén, M., Boström, G.: Security Engineering and eXtreme Programming: An Impossible Marriage? In: Zannier, C., Erdogmus, H., Lindstrom, L. (eds.) XP/Agile Universe 2004. LNCS, vol. 3134, p. 117. Springer, Heidelberg (2004)
20. Beck, K., et al.: Manifesto for Agile Software Development (February 2001)
21. The Agile Alliance Home Page, <http://www.agilealliance.org/home>

Accuracy Comparison of Predictive Algorithms of Data Mining: Application in Education Sector

Mamta Sharma and Monali Mavani

Faculty-MCA

SIES College of Management Studies

mamta_sharma1@hotmail.com,

monamavani@gmail.com

Abstract. Prediction is growing area of research which is attracting many researchers. Prediction is applied to almost all the sectors. Much commercial Business Intelligence software is available in which prediction is one of the features. With the advent of Open Source Technologies, it has become possible for education sector which normally has low IT budget, to take maximum advantage of Information and Communication Technologies (ICT). This paper describes the use of Open source Software Knime for prediction of students result based upon various independent (predictor) variables and value of dependent variable can be predicted using decision tree, SOTA (Self Organizing Tree Algorithm) and Naive Bayes. This paper compares these three predictive algorithms present in Knime in terms of accuracy. Predicted results are compared with the actual result in order to measure accuracy and recommends best Predictive algorithm for forecasting. This paper also demonstrates the use of Moodle - Open Source Learning Management System (LMS) Logs as one of the attributes in predicting the student results.

Keywords: Predictive algorithms, Decision tree, SOTA (Self Organizing Tree Algorithm), Naive Bayes, Knime, Moodle.

1 Introduction

Prediction is one of the techniques of data mining field. It has got high commercial value. With the growing competition among educational institutes, it is advantageous to extract some intelligent information from educational data. Educational data mining (EDM) is a field that exploits statistical, machine-learning, and data-mining algorithms over the different types of educational data. Its main objective is to analyze these types of data in order to resolve educational research issues [1]. The increase in both instrumental educational software as well as state databases of student's information have created large repositories of data reflecting how students learn [2]. On the other hand, the use of Internet in education has created a new context known as e-learning or web-based education in which large amounts of information about teaching-learning interaction are endlessly generated and ubiquitously available [3].

All this information provides a gold mine of educational data [4]. The Data Mining in education converts raw data coming from educational systems into useful information that could potentially have a great impact on educational research and practice and useful to management in certain decision making process.. This process does not differ much from other application areas of Data Mining , like business, genetics, medicine, etc., because it follows the same steps as the general Data Mining process [5]: preprocessing, Data Mining, and post processing. Prediction involves various data mining algorithms. Data Mining Algorithms can be classified in to supervised and unsupervised technique. This paper shows empirical comparison of three Data Mining algorithms used for prediction- Decision Tree, Naïve Bayes, SOTA (The Self-Organizing Tree Algorithm). We have used Open Source data mining software Knime for prediction and comparison. We also have used activity logs as one of the predictor variable which are obtained from LMS Moodle hosted in our college -SIES College Of Management Studies (SIESCOMS).

1.1 Supervised and Unsupervised Data Mining Technique

Methods for analyzing and modeling data can be divided into two groups: “supervised learning” and “unsupervised learning.” Supervised learning requires input data that has both predictor (independent) variables and a target (dependent) variable whose value is to be estimated. By various means, the process “learns” how to model (predict) the value of the target variable based on the predictor variables. Decision trees, regression analysis and neural networks are examples of supervised learning. Unsupervised learning does not identify a target (dependent) variable, but rather treats all of the variables equally. In this case, the goal is not to predict the value of a variable but rather to look for patterns, groupings or other ways to characterize the data that may lead to understanding of the way the data interrelates. Cluster analysis, correlation, factor analysis and statistical measures are examples of unsupervised learning. [6]

1.2 Decision Tree Classifier

A decision tree is a logical model represented as a binary tree that shows how the value of a *target variable* can be predicted by using the values of a set of input or *predictor variables*. The Decision Trees algorithm creates hierarchical structure of classification rules“ If ... Then ...”. To decide the branch nodes starting from the root, the questions like “Is the value of the parameter x greater than y?”. If the answer is positive, assign an object to right side branch, if it is negative assign an object to left side branch. A decision tree is constructed by a binary split that divides the rows in a node into two groups The same procedure is then used to split the subgroups. This process is called “recursive partitioning”. The split is selected to construct a tree that can be used to predict the value of the target variable. Decision tree learning is a common method used in data mining. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. [7]

1.3 SOTA Classifier

The Self-Organizing Tree Algorithm, (SOTA) (Dopazo, J. and Carazo, J.M. (1997) *J. Mol. Evol.*, 44, 226–233), is a neural network that grows adopting the topology of a binary tree. The result of the algorithm is a hierarchical cluster obtained with the accuracy and robustness of a neural network. SOTA clustering confers several advantages over classical hierarchical clustering methods. SOTA is a divisive method: the clustering process is performed from top to bottom, i.e. the highest hierarchical levels are resolved before going to the details of the lowest levels. The growing can be stopped at the desired hierarchical level. Growth of the tree can be stopped based on other criteria, like the allowed maximum Diversity within the cluster and so on. Moreover, a criterion to stop the growing of the tree, based on the approximate distribution of probability obtained by randomisation of the original data set, is provided. [8]

1.4 Navie Bayes Classifier

Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods. In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. [7]

1.5 Open Source Software MOODLE and KNIME (Konstanz Information Miner)

Moodle is a web-based, a free, open source software package Learning Management System (LMS), i.e. a Course Management System (CMS) and VLE designed around pedagogical principles, namely a social constructivist philosophy using the collaborative possibilities of the Internet. It is open to registered users and offers many different functions, ranging from course management to monitoring students' activities. It can be used as repository for course material, but it also offers the possibility to develop forums, wikis, quizzes, surveys, and other interactive inbuilt activities [9]. It has excellent documentation, strong support for security and administration, and is evolving towards Information Management System/Shareable Content Object Reference Model (IMS/SCORM) standards [10] [11][12]. It has great potential for creating a successful elearning experience by providing an abundance of excellent tools that can be used to enhance conventional classroom instruction in any VLE system. Moodle can scale from a single-teacher site to a more than 50-thousand-student University [13] Knime is a user-friendly and comprehensive Open-Source data integration, processing, analysis, and exploration platform. From day one, KNIME has been developed using rigorous software engineering practices and is currently being used actively by

over 6,000 professionals all over the world, both in industry and academia [14]. KNIME was developed (and will continue to be expanded) by the Chair for Bioinformatics and Information Mining at the University of Konstanz. The KNIME base version already incorporates over 100 processing nodes for data I/O, preprocessing and cleansing, modeling, analysis and data mining as well as various interactive views, such as scatter plots, parallel coordinates and others. KNIME allows users to visually create data flows (or pipelines), selectively execute some or all analysis steps, and later inspect the results, models, and interactive views. KNIME is written in Java and based on Eclipse and makes use of its extension mechanism to allow for plug-in providing additional functionality. Additional plug-in allow to add modules for text, image, and time series processing and the integration of various other open source projects, such as R (programming language), Weka, the Chemistry Development Kit, and LibSVM [15].

We have integrated these two software in order to achieve desired objective. In SI-ESCOMS Moodle is implemented since Jun 2009. Moodle collects all the information regarding users activities. These log information is taken and given to predictor model as one of the predictor variable. Knime's predictors above mentioned algorithms are used for prediction of students result which then compared with actual result. Accuracy is obtained in each case which is then compared to judge their performance.

2 Data Gathering and Analysis

Data of 120 students of MCA of SIES College of management studies was gathered from various sources like college records for results, survey data taken from students and students activity logs from moodle are combined together and given to data pre-processing stage. The various parameters in data are their Graduate degree, graduate result, their past semester result of MCA and their usage of Moodle was taken from Moodle logs. Various transformation techniques are applied on data to transform it in a form which knime's SOTA learner, Decision tree learner and Naive Bayes learner can understand. After getting learned models from these three algorithms, they are connected to their respective predictors as one of the input sources along with data to be predicted.

The requirements for a predictive model are as follows:

- **A single key column.** Dataset must contain one numeric or text column that Uniquely identifies each record. In our dataset roll no. acts as unique column
- **A predictable column.** Requires at least one predictable column. In our data set final result acts as predictable column.
- **Input columns.** Requires input columns, which can be discrete or continuous. Increasing the number of input attributes affects processing time but at the same time increases the efficiency.

Based on the above inputs Predictor is predicting students result. These predictions are validated against actual data to find out the accuracy of model as shown in Fig 1. The first icon in work flow diagram in Fig 1 is ARFF reader which reads data from ARFF (Attribute Relation File Format). After that data is passed to missing values icon which handles the missing values found in the data either by removing those

rows or by retaining them. From this point data is passed to all three predictive algorithms learners ie, SOTA learner, Naïve Bayes learner and Decision tree learner. All these three learners learn from the training data given to them and these learner models are used by their respective predictors to predict the data. Node Column comparator which compares the values of two columns (Predicted vs. Actual) and value counter which counts the number of occurrences of all values in a selected column were used to check the accuracy of each predictive algorithm.

It is found that accuracy of SOTA predictor is 76% , accuracy of decision tree is 98% and accuracy of naïve Bayes Predictor is 94% i.e. out of 120 records of students, 72 records were given to all learners and rest of the 48 records were used for predictor to predict. our data of students. In case of SOTA predictor 36 records had the predicted result same as actual result. In decision tree 47 records had the same result and in Naïve Bayes 45 records had similar values for actual and predicted result.

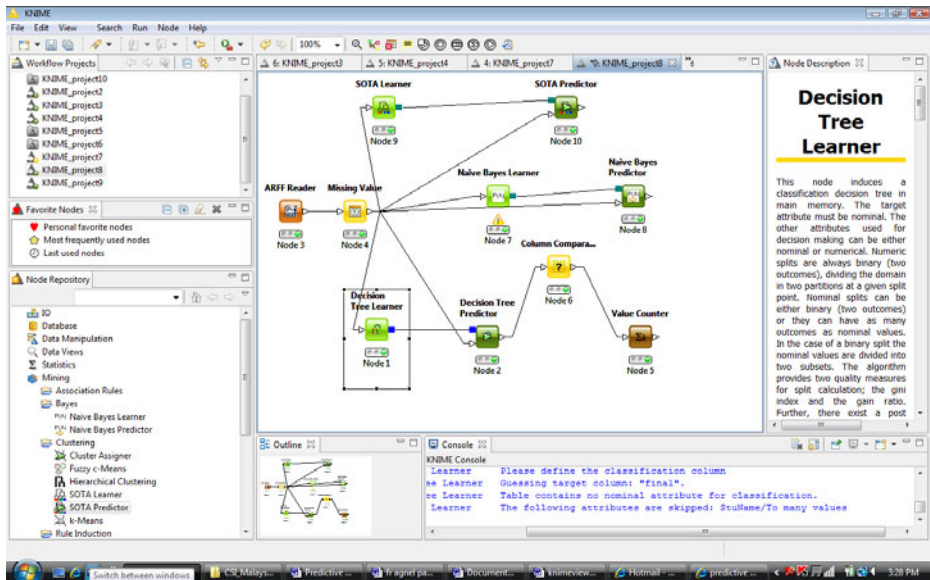


Fig. 1.

3 Conclusion and Future Work

The major advantage of predictive algorithms is that they can be easily interpreted. In this paper we have predicted only student's results. The predicted model says that if similar characteristics students come in the future then probability of their performance will be same as predicted. This helps educational institutes to adapt their methodologies in admission as well as in teaching like value addition in classes, teaching beyond the syllabus, student wise teaching methods etc. In this paper we have restricted to students data but that can be extended to various functional data of Education sector in the future. Many other Mining algorithms can be used to get more intelligence from education data.

References

1. Barnes, T., Desmarais, M., Romero, C., Ventura, S.: Presented at the 2nd Int. Conf. Educ. Data Mining, Cordoba, Spain (2009)
2. Koedinger, K., Cunningham, K., Skogsholm, A., Leber, B.: An open repository and analysis tools for fine-grained, longitudinal learner data. In: Proc. 1st Int. Conf. Educ. Data Mining, Montreal, QC, Canada, pp. 157–166 (2008)
3. Castro, F., Vellido, A., Nebot, A., Mugica, F.: Applying data mining techniques to e-learning problems. In: Jain, L.C., Tedman, R., Tedman, D. (eds.) *Evolution of Teaching and Learning Paradigms in Intelligent Environment. Studies in Computational Intelligence*, vol. 62, pp. 183–221. Springer, New York (2007)
4. Mostow, J., Beck, J.: Some useful tactics to modify, map and mine data from intelligent tutors. *J. Nat. Lang. Eng.* 12(2), 195–208 (2006)
5. Romero, C., Ventura, S., De Bra, P.: Knowledge discovery with genetic programming for providing feedback to course-ware author. *User Model. User-Adapted Interaction: J. Personalization Res.* 14(5), 425–464 (2004)
6. <http://www.dtreg.com>
7. <http://www.wikipedia.org>
8. Herrero, J., Valencia, A., Dopazo, J.: A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Oxford Journals, Bioinformatics* 17(2), 126–136 (2001)
9. Hong, R., Zhan, Y., Zhou, C.: A Knowledge Management System-based Model of E-learning for Higher Education in Chinese Context. In: *International Conference on Management and Service Science, MASS 2009*, pp. 1–4 (2009)
10. Berry, M.: An investigation of the effectiveness of Moodle in primary education, in Deputy Head, Haslemere (2005)
11. Zenha-Rela, M., Carvalho, R.: Work in Progress: Self Evaluation Through Monitored Peer Review Using the Moodle Platform. In: *36th Annual Frontiers in Education Conference. IEEE, San Diego* (2006)
12. Brandl, K.: Are you Ready to "Moodle"? In: *Language Learning/Technology*, Washington, vol. 9(2), pp. 16–23 (2005)
13. Al-Ajlan, S.A., Zedan, H.: Why Moodle. In: *12th IEEE International Workshop on Future Trends of Distributed Computing Systems*
14. Gonzalez-Barbone, V., Llamas-Nistal, M.: Trends in Content Reuse and Standardization. In: *Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference* (2007)
15. <http://www.knime.org>

Orchestrator Model for System Security

Aradhana Goutam¹, Rajkamal², and Maya Ingle²

¹ Fr. Conceicao Rodrigues College of Engineering, Bandstand, Bandra (W),
Mumbai 400050, Maharashtra, India

² School of Computer Science & IT, Devi Ahilya University, Indore 452001 MP, India
aradhana.pande@gmail.com, dr_rajkamal@hotmail.com,
maya_ingle@rediffmail.com

Abstract. Service Oriented Architecture recommends a better flexibility & increase efficiency by reusing the services. But it also increases the complexity w.r.t. Security. In today's era Orchestration is basic requirement for the service based security. Orchestration is a key control mechanism that invokes Services to work, as well as to provide control. In this paper we are going to represent the service interface for security constraints which will make their composition safe. Main aim is to the increase reusability and quickness, bring down difficulty, increase advanced extensibility and improved scalability and reliability.

In this paper we are introducing a new model which consists of Services, security and Orchestration together to improve the Security services.

Keywords: Orchestrator, Security Services.

1 Introduction

From last few years IT environment is changing rapidly. According to business needs many organizations are aiming to achieve flexible system. SOA offers the organizations to exchange the Data, Partners and their services across the companies. For this kind of communication secured environment is required. We are aiming to find a better way for high reusability of services with security without impacting organization's business.

Some specific approaches expose the best practices for securing SOAs. Most of them focus on the development of the Enterprise Service Bus (ESB) [1,2] or similar middleware solutions [4]. In this paper we are introducing the *Orchestration model for System Security (OSS)*. This model consist of Security services and central controlling unit using orchestration.

1.1 General Security Architecture

General security architecture is shown in Figure 1. The security architecture has two parts—:

Public Network-- This part of architecture is public web world.

Private Network-- This part contains the Private web area of any organization or firm. In this area there are different security related services. This area is sheltered or

protected web area. Above both the parts are connected to each other through an end point which is controlling the communication between public network and private network area.

Advantages of this architecture are it will decrease complexity and development cost, higher alertness.

The drawback with the service based architecture is Interoperability troubles, decentralized routing (each service), lower performances.

This drawback can be eliminated through an Orchestration based security system.

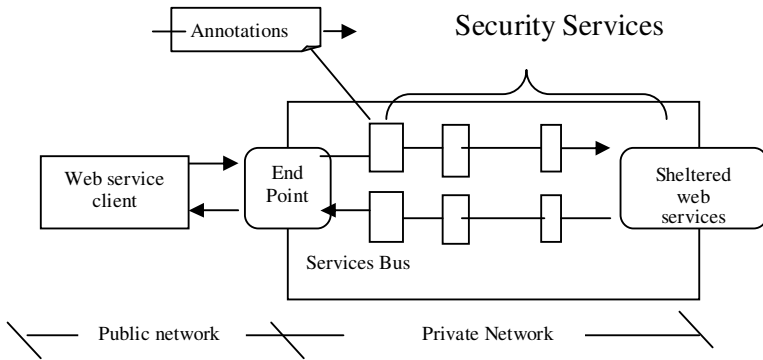


Fig. 1. Service based security architecture

2 Motivation

A key objective of SOA consists of increasing the flexibility and the agility of the business [3]. One of the objective is to introduce new security component. Second objective is to provide security according to the requirement of the organization. This approach is based on Aspect Oriented Programming (AOP).

3 Security Services

These services help to the SOA for optimized and simple services. Security services are centralized which can be used from the server. This will help for improving flexibility and efficiency of the system. A main instance for Security services is orchestrations that propose service composition. Security Services will offer Services with Security. This will increase the reusability and decreases the complexity.

Security Services classification-

Authentication: This will check the validity like digital certificate, digital signature etc.

Authorization: This will check the significant information, policies etc.

Gateway: This is the interface between the incoming and outgoing information.

Verification & Validation: These will ensures incoming and outgoing information.

Filtering: This will do the prevention from the hijacking, Denial of Services attack and buffer overflow and etc

Cryptography: This is used for the Encryption and Decryption of incoming and outgoing packets.

4 Proposed Architecture

The Proposed architecture contains Orchestration, security services in trusted area. Any incoming message will be passed through the gateway and handed over to the Orchestrator. According to the response message Orchestrator will invokes the Security services.

If any check failed than an exception will be thrown to the orchestrator and it will stop services to the user. Figure 2 shows the use case, as user wants to access service 2 and require security means (step 0). This will invoke public registry and message will be generated. At next step Gateway and orchestrator Works as a single entity and will capture message (step 1). Now it will search for the Protected Policy Services. The Orchestrator will invoke services like Authentication, validation, Authorization. Now all the services work in order to 2, 3, 4, 5, 6, 7. if whole process passed then Orchestrator will respond to user through step 9, and if any of the above step fails then orchestrator will stop processing and send a deny message to the user.

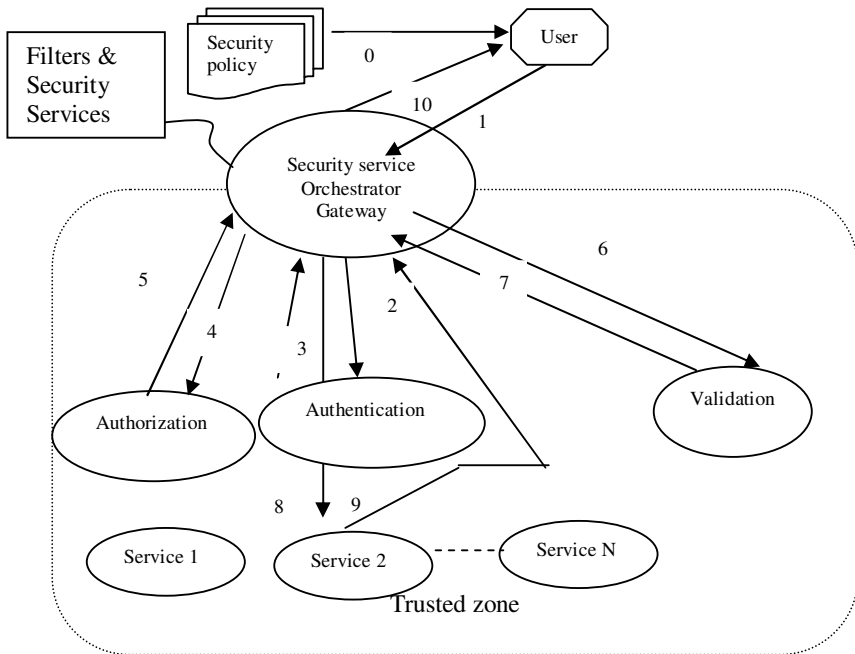


Fig. 2. Use Case

Orchestrator can be used in ways:

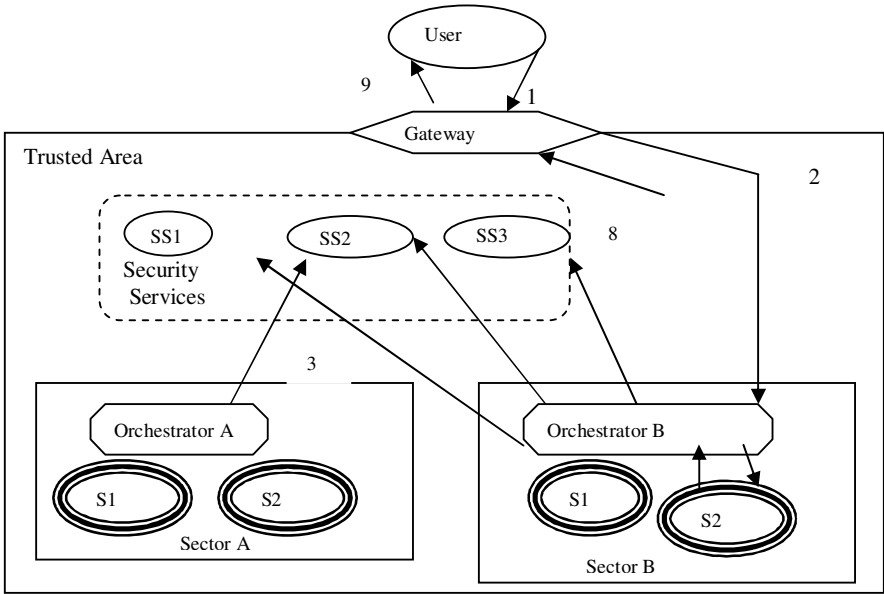


Fig. 3. Per Policy Orchestrator

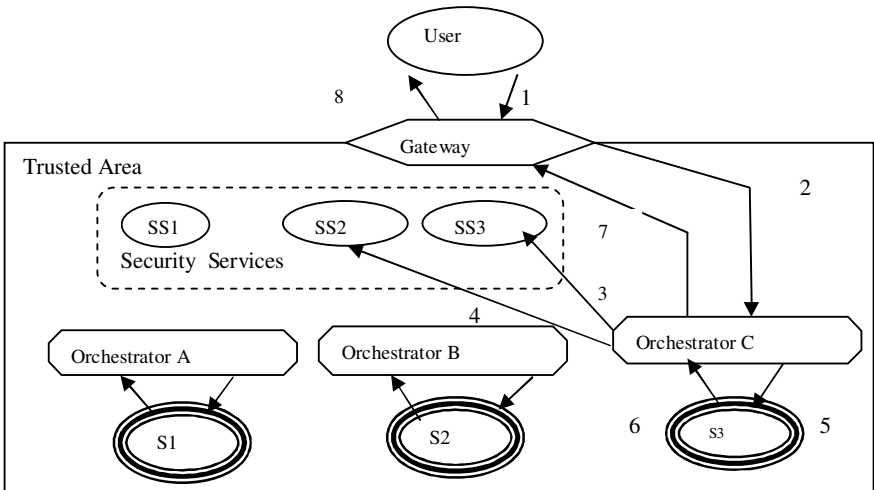


Fig. 4. Per Service Orchestrator

4.1 Per Policy Orchestrator

The first approach one dedicated orchestrator is assign for single group of services required same security functionality. In this way each orchestrator is responsible for its own functionality ands services. As in Fig. 3 SS1, SS2, SS3 are the Security

Services, and Each sector have its own services i.e. s1,s2,s3 and so on. The orchestrator of sector A is using only one Security Service i.e. SS2, but Orchestrator of Sector B is using All three services SS1, SS2, SS3. The decomposition into Sectors will distribute the load on several orchestrator. But this will increase the complexity of architecture because for each group of services one orchestrator is required.

4.2 Per Service Orchestrator

The second approach implements one orchestrator for one service. In this manner change in policy will change in orchestrator also. An Orchestrator for one Service can avoid the linking compatibility problem.

5 Conclusion

In this paper, we proposed a new architectural framework for security systems called as Orchestrator Model for System Security (OSS). The architecture is contains three components: services, security services and orchestrator.

This results to the separation of the security aspect from business services which reduce design complexity and development costs. In addition, orchestration solution centralize the security system and therefore easier to control and administrate. This results into the separation of security services with services which reduces the development cost and Complexity. Orchestrator working as centralized solution for security therefore it's easy to administrate and control in centralized manner.

References

- [1] Chappel, D.: Enterprise Service Bus (2004)
- [2] Hohpe, G., Woolf, B.: Enterprise Integration Patterns, Designing, Building, and Deploying Messaging Solutions (2004)
- [3] Report. State of soa adoption report – gauging the use of soa systems in the enterprise (January 2008)
- [4] Khalaf, R., Nagy, W., Curbera, F., Duftler, M.J.: Colombo: Lightweight middleware for service-oriented computing. IBM Journal of Research and Development 44(4) (2005)
- [5] Montesi, F., Guidi, C., Lucchi, R., Zavattaro, G.: JOLIE: a java orchestration language interpreter engine. Electronics Notes in Theoretical Computer Science 181, 19–33 (2007)
- [6] Types and Effects for Secure Service Orchestration
- [7] IBM Tivoli Staff, <http://www.ibm.com/developerworks/tivoli/library/t-tiocitrix/index.html>
- [8] Bartoletti, M., Degano, P., Ferrari, G.L.: Types and Effects for Secure Service Orchestration. In: Proceedings of the 19th IEEE workshop on Computer Security Foundations (2006) ISBN ~ ISSN:1063-6900
- [9] Busi, N., Gorrieri, R., Guidi, C., Lucchi, R., Zavattaro, G.: Choreography and Orchestration conformance for system design. In: Ciancarini, P., Wiklicky, H. (eds.) COORDINATION 2006. LNCS, vol. 4038, pp. 63–81. Springer, Heidelberg (2006), doi:10.1007/11767954_5

Performance Analysis of Interleave Division Multiple Access Scheme with Different Coding Techniques

Parul Awasthi, Sarita Singh Bhadauria, and Madhuri Mishra

Abstract. This paper gives the performance analysis of Interleave Division Multiple Access (IDMA) with different coding techniques. Convolutional coding and Zigzag coding techniques are used to improve the capacity of a channel. A new multiple access scheme has been introduced which employs the chip level interleaving as the only means of user separation. Coded IDMA is compared with un coded IDMA in this paper.

Keywords: CDMA, IDMA, Interleavers, zigzag code, convolutional code.

1 Introduction

First and second generation cellular systems are dominated by orthogonal MA approaches. The main advantage of these approaches is the avoidance of intra-cell interference. However, careful cell planning is necessary in these systems to curtail cross-cell interference. In particular, sufficient distance must exist between re-used channels, resulting in reduced cellular spectral efficiency. Non-orthogonal CDMA techniques have been adopted in second and third generation cellular systems (e.g, IS-95, CDMA2000 and uplink WCDMA). Compared with its orthogonal counterparts, CDMA is more robust against fading and cross-cell interference, but is prone to intracell interference. Due to its spread- spectrum nature, CDMA is inconvenient for data services. The main advantages of CDMA are its robustness against fading and cross-cell interference and its flexibility in asynchronous transmission environments. A main concern for CDMA is intracell interference. This can be treated by MUD [7] however, the application of MUD has been limited by its high computational cost (complexity may be increased in an exponential or polynomial order with K , the number of users involved). It is also difficult to support high single-user rate with CDMA.

With recent progress in iterative detection technique, MUD complexity can be reduced to a level comparable to single-user detection for orthogonal schemes. A potential scheme is IDMA that relies on the user-specific interleavers for user separation. The spreading operations in CDMA can be replaced by low rate forward error correction (FEC) codes in IDMA to provide increased coding gain. Besides cost issues, there are also theoretical ones. Although the existing RW-CDMA systems (such as IS-95) have advantages in multi-cell cellular environments, they have relatively low throughput in single-cell environment compared with FDMA and TDMA. This is mainly due to the effect of the same-cell user MAI. However, the fundamental work in [6] shows that, at least theoretical, there is no penalty on

throughput with RW-CDMA even in single-cell environments, provided that the signaling method is designed properly.

Let R be the rate of the FEC code C and N the length of the signature sequence used by the spreaders. Assume the same R and N for all the users. After multiple accesses by the K users, on an average $R \times K/N$ information bits are transmitted across the channel during chip duration, so we refer to $R \times K/N$ as the system throughput. We denote by η the maximum achievable value $R \times K/N$ for a given R for reliable communication (η is called the spectral efficiency).

It is seen from Fig.1 that η is maximized when R is minimized for any fixed E_b/N_0 . Some consequences are as follows:

- Assume that we fix the bandwidth expansion factor (N/R) for each user. Assume that we fix the bandwidth expansion factor (N/R) for each user. The minimizing R implies minimizing N . The minimum value of N is 1, meaning no spreading.
- RW-CDMA incurs no penalty on throughput when $R \rightarrow 0$, since it can be verified that η approaches the unconstrained AWGN channel capacity when $R \rightarrow 0$
- There will be a non-negligible penalty on system throughput if R is not sufficiently small.

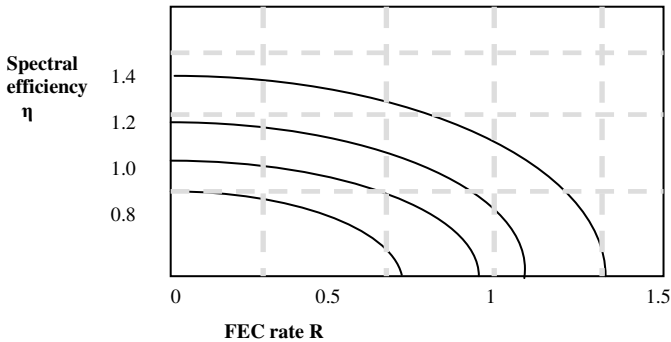


Fig. 1. The η versus R curves for an RW-CDMA system with AWGN and $K \rightarrow \infty$

As a special case of CDMA, IDMA inherits most of the advantages of CDMA. The performance of some illustrative IDMA system is shown in fig.2. A conventional random waveform CDMA (RW-CDMA) system (such as IS-95) involves separate coding and spreading operations. The optimal multiple access channel (MAC) capacity is achievable only when the entire bandwidth expansion is devoted to coding. This suggests combining the coding and spreading operations using low rate codes to maximize coding gain. However, how can we separate different users without spreading within the CDMA framework?

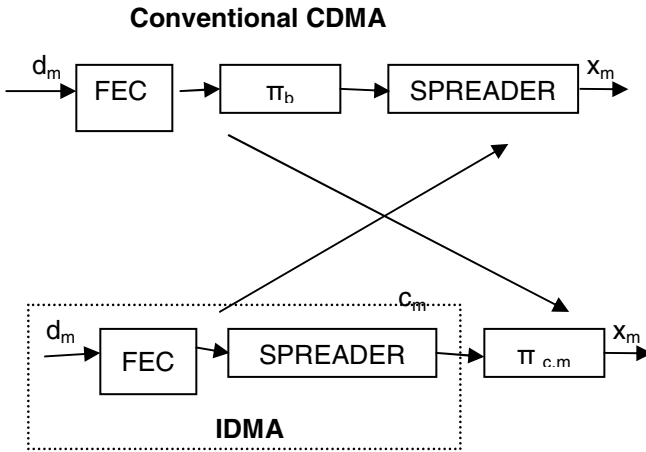


Fig. 2. Illustration of IDMA system

One possible solution of this problem is to employ chip level interleavers [3] for user separation. This scheme is a special case of CDMA in which bandwidth expansion is entirely performed by low rate coding. IDMA inherits many advantages of CDMA, in particular, diversity against fading and mitigation of the worst case other cell user interference problem.

2 System Description

2.1 Transmitter Structure

In conventional CDMA scheme data of user k is encoded by a forward error correction (FEC) code followed by an interleaver π_k . A spreading operation is then applied to produce the transmitted signal. A conventional CDMA scheme, such as IS-95, employs FEC coding rate at around $1/2 \sim 1/3$ [1]. Quite long signature sequences are required to support a large number of users. This is not an optimized approach (although perhaps a convenient one) according to the above discussion. Clearly, some structural change is necessary if we want to make progress towards fully exploiting the available capacity. A key issue is to how to achieve multiple accesses in low-rate coded systems. In order to simplify EMUD and to minimize the FEC code rate IDMA transmitter is employed. The key principle of IDMA is that the interleavers $\{ \pi_k \}$ should be different for different users. We assume that the interleavers are generated independently and randomly. These interleavers disperse the coded sequences so that the adjacent chips are approximately uncorrelated, which facilitates the simple chip-by-chip detection scheme.

Fig. 3 shows the transmitter structure of the multiple access schemes under consideration with K simultaneous users. The input data sequence \mathbf{d}^k of user- k is encoded based on a low-rate code C , generating a coded sequence $\mathbf{c}^k!$ $[\mathbf{c}^k(1), \dots, \mathbf{c}^k(j), \dots, \mathbf{c}^k(J)]^T$, where J the frame length. The elements in \mathbf{c}^k are referred to as

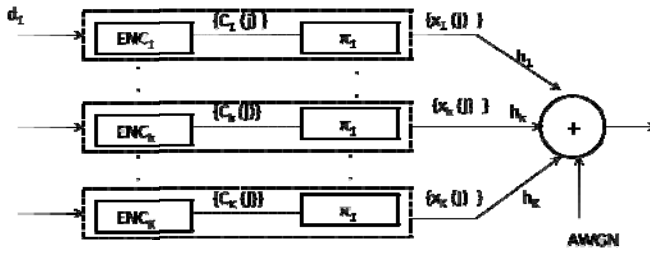


Fig. 3. Transmitter structure of the multiple access scheme under consideration with K simultaneous users

coded bits. Then ck is permuted by an interleaver k , producing $xk = [xk(1), \dots, xk(j), \dots, xk(J)]^T$. Following the CDMA convention, we call the elements in xk “chips”. Users are solely distinguished by their interleavers, hence the name interleave-division multiple-access (IDMA).

2.2 Receiver Structure

In conventional CDMA, at the receiver side, the received signal is passed through a bank of correlators. A turbo process is then applied involving two functions: an elementary multi user detector (EMUD) and a bank of K soft in soft out decoders (DECs) based on C . In the receiver (fig 4.), two constraints must be considered (1) the FEC code C and (2) the correlation among signature sequences. Finding a joint optimal solution is usually computationally prohibitive. The turbo processor takes a sub optimal approach by decomposing the task in to two parts. The EMUD processes constraint the correlation among signature sequences and ignores the FEC code C . The DECs process the FEC code C and ignores the correlation among signature sequences. A global iterative process is then applied to refine the results.

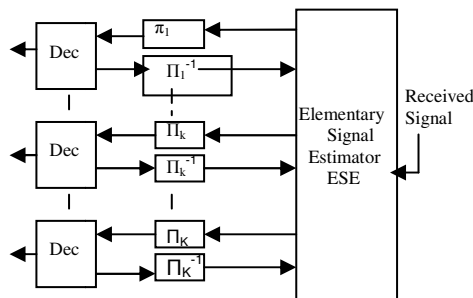


Fig. 4. Receiver structure

The DEC complexity per user is independent of the simultaneous user number K . The EMUD complexity per user, on the other hand, increases rapidly with K . This is a concern when K is large. Hence to minimize the EMUD cost without seriously

affecting the DEC cost a potential method is to move spreading operation into FEC coding so as to reduce the EMUD cost related to spreading. This is the strategy to be in IDMA receiver.

The receiver operation is still based on two constraints

(1) The constraint of the FEC code C and (2) The constraint due to the superposition of the transmitted chips.

Without signature sequences, the processing related to the second constraint becomes very simple. The chip interleavers allow adopting a chip – by-chip estimation technique [3]. Flowchart of decoding algorithm is presented in Fig.5. Number of users and iteration value is initialized and eDEC is set to zero. Decoding results will become better by increasing the number of iterations.

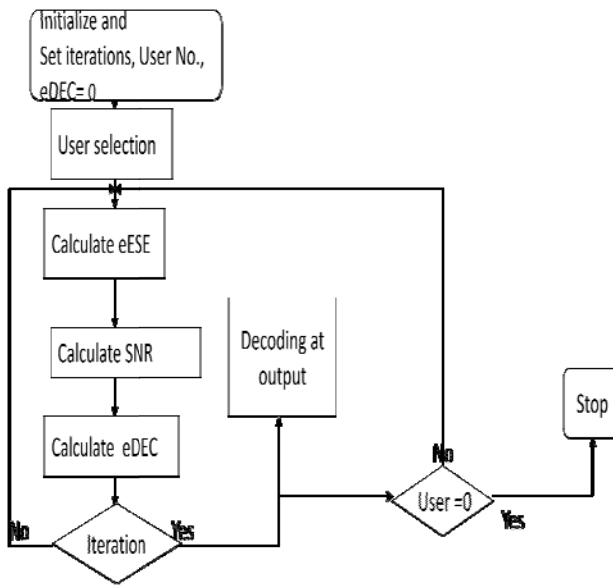


Fig. 5. Flowchart of decoding algorithm

3 Analysis and Simulation Results

3.1 For Convolutional Coded IDMA System

In this section, we examine the impact of interleaver on system performance. In fig.6 for 32 users with spreader length 16 and block 20 (each block contains 1024 bits) convolutionally coded system with code rate 1/2 and un coded system is compared. With convolutional coding results are much better than without any coding scheme. For signal to noise ratio of 3 dB, bit error rate is same for both. If we make SNR value double bit error rate is reduced by a factor of 10 approximately and if SNR is 9 (thrice) then it is reduced approximately by a factor of 100.

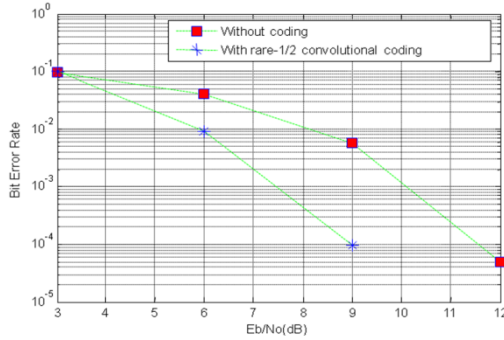


Fig. 6. Comparison of convolutional coded system with uncoded system

3.2 For Zigzag Coded IDMA System

Zigzag code [8] can be considered as a two-state rate systematic convolutional code with generator polynomial matrix $G(X) = [1 \quad 1 / 1+X]$. The parity sequence is punctured so that only one parity bit is transmitted for every J information bits—thus, reducing the rate to $J / J + 1$. The trellis-based APP and MLA decoders of the convolutional code, however, are much less efficient than the alternative algorithms based on the zigzag representation.

In fig. 7 for 16 users zigzag coded system is compared with uncoded system with same simulation parameters and it is found that it is much better than un coded IDMA System for higher values of signal to noise ratios.

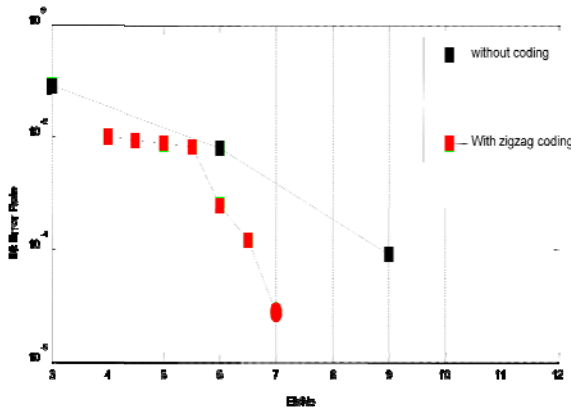


Fig. 7. Comparison of zigzag coded system with uncoded system

4 Interleaving Schemes

The principle of traditional periodic interleaving scheme which is suitable to block codes can be expressed as:

Let the interleaving degree is I . At first, $I(n, k, t)$ linear block codes are arranged in rows in an array $I \times n$. Then we transmit the array column by column. And at the receiver, the received data are rearranged in the same array column by column, then decoding rank by rank. That is the whole interleaving procedure. Because the code length n of block code which has better error correcting performance is often long and $(n - k)$ check bits of each code are only related to k information bits, without regard to other codes, interleaving technique can separate long burst errors effectively to different codes. But n is often small for (n, k, N) convolutional code. $N \times n$ bits are related to each other in coding and decoding procedure. So, the above interleaving scheme can't separate effectively long burst errors that still lie in relevant codes and decoding will be not satisfactory the following is to investigate two interleaving schemes suitable to convolutional codes.

1. Block Interleaving Scheme

Considering performance improvement and time delay, 1000 codes after encoding can be seen as a block to interleave. If the interleaving degree is I , the interleaving procedure can be described as:

At first, we transmit the code in 1000 codes. Then transmit the $(I+1)$ th code, then $(2I+1)$ th code, until $xI+1 > 1000$ ($x = \text{int}(1000/I)$ or $x = \text{int}(1000/I) - I$) after then, the second code will be sent. Then the $(I+2)$ th code, the $(2I+2)$ th code, until $xI+2 > 1000$. We transmit 1000 codes just like the above. The next block with 1000 codes will be sent also this way. At the receiver, make a reverse work on each block with 1000 codes, which is deinterleaving. The advantage of this interleaving scheme is that it can separate long burst errors effectively to irrelevant codes and avoid "error propagation" in decoding procedure when $I > N$.

2. Bit interleaving scheme

Long burst errors can be separated to irrelevant codes using block interleaving scheme, but may be still lie in n bits of one code. To separate long burst errors more effectively, taking $(2, 1, 3)$ convolutional code with interleaving degree I .

Step 1: 1000 $(2, 1, 3)$ convolutional codes after encoding are seen as a block to interleave.

Step 2: The first bit of the first code 1000 codes is sent into the channel firstly. Then the first bit of $(I+1)$ th code, then the first bit of the $(2I+1)$ th code, until the first bit of the $(xI+1)$ th code is transmitted.

Step 3: Afterwards, the second bit the first code will be transmitted. Then the second bit of the $(I+1)$ th code, then the second bit of the $(2I+1)$ th code, until the second bit of the $(xI+1)$ th code is transmitted.

Step 4: The first bit of the second code is transmitted afterwards, then the first bit of the $(I+2)$ th code, then the first bit of the $(2I+2)$ th code, until the first bit of the $(xI+2)$ th code is sent into the channel.

Step 5: Circulating like the above, 2000 bits of 1000 codes are all sent into the channel.

Step 6: Go back to step 1, the next block with 1000 codes will be interleaved.

5 Conclusion

In this paper, an insight is initially given regarding the transmitter and the receiver structure for IDMA. We provide an approximate analysis for the random interleavers. It was further used in the scheme to track the effect of bit error ratio as the signal to noise ratio increases. Two different coding schemes i.e. convolutional coding and zigzag coding are discussed and both are applied to IDMA system. We analyze the system for both the coding techniques with same simulation parameters of an uncoded system. Coded system gives much better result than uncoded system. The reason for applying the two coding schemes is to find out which particular scheme is much better for system. As a result we found that if signal to noise ratio is incremented by same value as of SNR the bit error rate is reduced by a factor of 10 approximately each time. Random interleavers required much memory to store the interleaving sequence. So if by using some other interleaver such as Master interleaver memory requirement is reduced then coded system will give much better results. Hence by our thesis we have explained the feasibility and advantages of coded interleaver based multiple access scheme together with an accurate and effective prediction technique.

IDMA inherits many advantages from CDMA, in particular, diversity against fading and mitigation of the worst-case other-cell user interference problem. It has also been shown that the performance of IDMA is much better than that of CDMA as the number of users is increased. It has also explained the feasibility and advantages of the interleaver-based multiple access schemes together with an accurate and effective performance prediction technique. We expect that the basic principles can be extended to other applications, such as space-time codes [4] and ultra wideband (UWB) systems, MIMO systems, a hybrid scheme inheriting the advantages non-orthogonal scheme, OFDM-IDMA.

References

1. Ping, L., Liu, L., Wu, K., Leung, W.: Interleave Division Multiple Access. IEEE Transactions on Wireless Communications 5(4), 938–947 (2006)
2. Bie, H., Bie, Z.: A Hybrid multiple Scheme. IEEE Transactions on Wireless Communications (October 2006)
3. Liu, L., Leung, W.K., Ping, L.: Simple chip-by-chip multi-user detection for CDMA systems. In: Proc. IEEE VTC 2003-Spring, Jeju, Korea, pp. 2157–2161 (April 2003)
4. Ping, L., Liu, L., Wu, K., Leung, W.: A unified approach to multi user detection and space time coding with low complexity and nearly optimal performance. In: Proc. 40th Allerton Conference, Allerton House, pp. 170–179 (October 2002)
5. Fishler, E., Poor, H.V.: On the tradeoff between two types of processing gain. In: Proc. 40th Allerton Conference, Allerton House, pp. 1178–1187 (October 2002)
6. Verdú, S., Shamai, S.: Spectral efficiency of CDMA with random spreading. IEEE Trans. Inform. Theory 45, 622–640 (1999)
7. Moher, M.: An iterative multiuser decoder for near-capacity communications. IEEE Trans. Commun. 46, 870–880 (1998)
8. Li, K., Yue, G., Wang, X., Ping, L.: Low-Rate Repeat-Zigzag-Hadamard Codes. IEEE Transactions on Information Theory 54(2) (February 2008)

Channel Assignment to Minimize Interference in Multiradio Wireless Mesh Networks

S. Sekhar Babu¹ and V. Sumalatha²

¹ JNTUA College of Engineering, Anantapur, Andhra Pradesh
sekhar_sappidi_01@yahoo.co.in

² Dept. of ECE, JNTUA College of Engineering, Anantapur, Andhra Pradesh
sumaatp@gmail.com

Abstract. Breadth First Search Channel Assignment(BFSCA) is a hybrid channel assignment algorithm that utilize multiple radio interfaces to improve the throughput and minimize the interference within the wireless mesh network and between the mesh network and co-located wireless mesh networks. This new channel assignment scheme allow different nodes in the same network to communicate with each other without causing too much interference to their neighbors. It is introducing Multiradio Conflict Graph(MCG) to model interference in the wireless mesh network. Breadth First Search Channel Assignment considers both the fixed channels(static) and the dynamic channels to reduce interference of the network. BFSCA will increase the network throughput greatly.

Keywords: Multiradio wireless mesh networks, channel assignment, Interference.

1 Introduction

Wireless mesh networks [1] are multihop networks of wireless routers. There is an increasing interest in using wireless mesh networks as broadband backbone networks to provide network connectivity in enterprises, campuses and in metropolitan areas. An important design goal for wireless mesh networks is capacity. It is well known that wireless interference severely limits network capacity in multihop settings [2]. One common technique used to improve overall network capacity is use of multiple channels [3]. Presence of multiple channels requires us to address the problem of which channel to use for a particular transmission; the overall objective of such an assignment strategy is to minimize the overall network interference. By deploying multi-radio routers in wireless mesh networks and assigning the radios to non-overlapping channels, the routers can communicate simultaneously with minimal interference in spite of being in direct interference range of each other. Therefore, the capacity of wireless mesh networks can be increased.

Problem Addressed. In our article, we address the problem of static and dynamic assignment of channels to links in the context of networks with multi-radio nodes [5]. The objective of the channel assignment is to minimize the overall network interference. Channel assignment is done using hybrid channel assignment algorithm “Breadth First Search Channel Assignment”. The assignment of channels to links must

obey the interface constraint that the number of different channels assigned to the links incident on a node is at most the number of interfaces on that node. To exploit the interference of nodes as well as radios, it uses multi-radio conflict graph (MCG).

The salient features of our work that set us apart from the existing channel assignment approaches on multi-radio platforms are as follows.

Our approach is “topology preserving,” i.e., all links that can exist in a single channel network also exist in the multi channel network after channel assignment. Thus, our channel assignment does not have any impact on routing.

2 Problem Formulation

The channel assignment algorithm we propose in this paper is designed for wireless mesh networks. Routers in such networks are stationary. However, user devices, such as laptops and PDAs, can be mobile. Such devices associate with routers that also function as access points. Fig.1 illustrates our model of a multi-radio mesh network. The routers need not all be equipped with the same number of radios nor do they need all three types of radios. Depending on the number of radios at each mesh router, we classify the routers into two categories: (1) Multi-Radio mesh routers (MRs); and (2) Single-Radio mesh routers (SRs). We mandate that each MR and SR in the network be equipped with one radio, called the default radio, which is of the same physical layer type, e.g. 802.11b, and tuned to the same channel. At least one router in the mesh is designated as a gateway. The gateways provides connectivity to an external network. In order to simplify the explanation of the channel assignment solution, we assume the presence of only one gateway. Access Points (APs) provide connectivity to user devices and are co-located with mesh routers. A majority of the traffic within the mesh is either from the user devices to the gateway or vice-versa. Therefore, in order to improve overall network capacity, it is preferable to place MRs close to the gateway and in regions of the mesh that are likely to experience heavy utilization. The dotted lines in fig.1 illustrate links between Mrs that are tuned to non-overlapping channels.

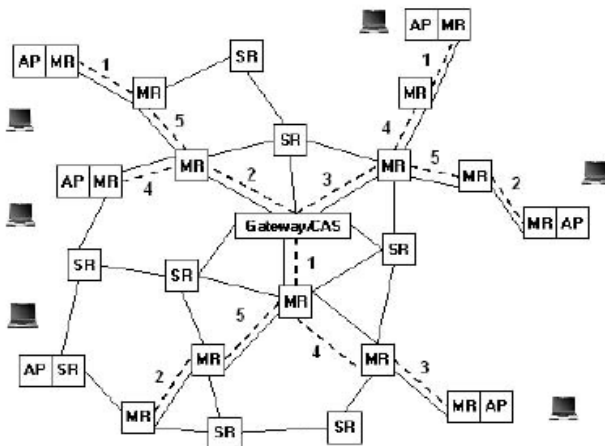


Fig. 1. Multiradio wireless mesh Architecture

A. Interference Estimation

The goal of interference estimation is to periodically measure the interference level in each mesh router’s environment.

An interfering radio is defined as a simultaneously operating radio that is visible to a router but external to the mesh. The interference estimation procedure is as follows: a mesh router configures one radio of each supported physical layer type to capture packets on each supported channel for a small duration. The router uses the captured packets to measure the number of interfering radios and per second channel utilization. The number of interfering radios is simply the number of unique MACs external to the mesh. The utilization on each channel due to the interfering radios is computed from the captured data frames by taking into account the packet sizes and the rates at which the packets were sent [4].

Each mesh router then derives two separate channel rankings. The first ranking is according to increasing number of interfering radios. The second ranking is according to increasing channel utilization. The mesh router then merges the rankings by taking the average of the individual ranks. The resulting ranking is sent to the CAS.

B. Interference Modeling: Multiradio Conflict Graph

Interference Modeling Conflict graphs are used extensively to model interference in cellular radio networks. A conflict graph for a mesh network is defined as follows:

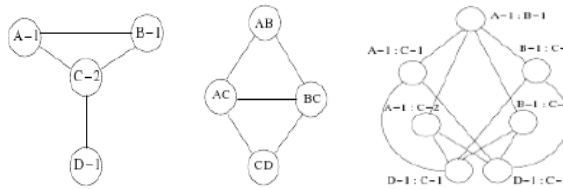


Fig. 2. (a) Network Topology (b) Conflict Graph(F) (c) Multiradio Conflict Graph F’

consider a graph, G, with nodes corresponding to routers in the mesh and edges between the nodes corresponding to the wireless links. A conflict graph, F, has vertices corresponding to the links in G and has an edge between two vertices in F if and only if the links in G denoted by the two vertices in F interfere with each other. As an example of a conflict graph, Fig. 2.(a) shows the topology of a network with four nodes. Each node in the figure is labeled with its node name and its number of radios. Fig. 2.(b) shows the conflict graph.

The conflict graph does not correctly model routers equipped with multiple radios. Therefore, is extended the conflict graph to model multiradio routers. In the extended model, called the Multi-radio Conflict Graph (MCG), are presented edges between the mesh radios as vertices instead of representing edges between the mesh routers as vertices as in the original conflict graph. To create the MCG, F', we first represent each radio in the mesh network as a vertex in G' instead of representing routers by vertices as in G. As an example, Fig. 2(c) shows the multiradio conflict graph of the network.

3 Breadth First Search Channel Assignment Algorithm

A. Overview

The proposed system uses a breadth first search to assign channels to the mesh radios. The search begins with links emanating from the gateway node. Before using the BFS-CA algorithm, the channel assignment server (CAS) obtains the interference estimates from the mesh routers. It then chooses a channel for the default radios. The default channel is chosen such that its use in the mesh network minimizes interference between the mesh network and co-located wireless networks. The CAS then creates the MCG for the non-default radios in the mesh. The MCG is created using the neighbor information sent by each mesh router to the CAS. After constructing the MCG, the CAS uses the BFS-CA algorithm to select channels for the non-default radios. Once the channels are selected for the mesh radios, the CAS instructs the routers to configure their radios to the newly selected channels.

B. Default Channel Selection

The CAS chooses the default channel using the rank of a channel, c , for the entire mesh, R_c . R_c is computed as follows:

$$R_c = \frac{\sum_{i=1}^n Rank_c^i}{n} \quad (1)$$

where n is the number of routers in the mesh and $Rank_c^i$ is the rank of channel c at router i . The default channel is then chosen as the channel with the least R_c value. Using such a channel satisfies our goal of minimizing interference between the mesh and co-located wireless networks.

C. Non-default Channel Selection

In this phase, the CAS uses the neighbor information collected from all routers to construct the MCG. Neighbor information sent by a router contains the identity of its neighbors, delay to each neighbor, and interference estimates for all channels supported by the router's radios. The CAS also associates with each vertex a channel ranking derived by taking the average of the individual channel rankings of the two radios that make up the vertex. For all vertices in the MCG, the CAS then computes their distances from the gateway. The distance of an MCG vertex is the average of the distances from the gateway of the two radios that make up the vertex. The distance of a radio is obtained from beacons initiated by the gateway. A beacon is a gateway advertisement broadcast hop-by-hop throughout the mesh.

Algorithm 1. BFS-CA Algorithm

- 1: Let $V = \{v | v \in MCG\}$
- 2: while notAllVerticesVisited $\{V\}$ do
- 3: Let $h = \text{smallestHopCount}(V)$
- 4: $Q = \{v | v \in V \text{ and } \text{notVisited}(v) \text{ and } \text{hopcount}(v) == h\}$
- 5: sort(Q)

```

6: while size(Q) > 0 do
7: vcurrent = removeHead(Q)
8: if visited(vcurrent ) then
9: continue
10: end if
11: visit(vcurrent )
12: Vn = {u | u ∈ MCG and edgeInMCG(u, vcurrent ) == TRUE}
13: permanently assign highest ranked channel c from vcurrent 's channel
ranking that does not conflict with ui , {ui ∈ Vn and 0 ≤ i < size(Vn )}
14: if c does not exist then
15: permanently assign random channel to vcurrent
16: end if
17: L = {v | v ∈ MCG and v contains either radio from vcurrent }
18: removeVerticesInListFromMCG(L)
19: tentatively assign c to radios in L that are not part of vcurrent
20: Let rf be router with interface in vcurrent that is farthest away from gateway
21: Let Tail = list of all active v (v ∈ MCG) such that v contains an interface
from rf
22: sort(Tail)
23: addToQueue(Q, Tail)
24: end while
25: permanently assign channels to radios that are unassigned a permanent
channel.
26: end while

```

Algorithm: The algorithm starts by adding all vertices from the MCG to a list, V (Line 1). In line 3, the smallest hopcount vertex is determined of all vertices in the MCG. In line 4, all vertices with distance equal to the smallest hop count are added to a queue, Q. These vertices are then sorted by increasing delay values (Line 5). This sort is performed in order to give higher priority to the better links emanating from the shortest hop count router (the gateway for the first BFS iteration). The algorithm then visits each vertex in Q (Line 11) and permanently assigns them the highest ranked channel that does not conflict with the channel assignments of its neighbors (Line 13). Once a vertex is assigned a channel, all vertices that contain either radio from the just-assigned vertex are placed in a list, L (Line 17). In line 18, all vertices from L are removed from the MCG. This step is needed to satisfy the constraint that only one channel is assigned to each radio. The radios in the list of vertices that do not belong to the just-assigned vertex are tentatively assigned the latter's channel (Lin 19). In lines 20-21, vertices at the next level of the breadth first search are added to Q. These vertices correspond to links that fan-out from the gateway towards the periphery. To find such links in the MCG, two steps are performed. In the first step (Line 20), the router from the just-assigned vertex that is farthest away from the gateway is chosen; the farthest router is the router with the higher hop-count of the two routers that make up the just-assigned vertex. In the second step (Line 21), all unvisited MCG vertices that contain a radio belonging to the farthest router are added to the list, Tail. This list

is sorted (Line 22) by increasing value of the delay metric to give higher priority to better links that emanate from the farthest router. Finally, in line 23, the vertices from T are added to Q .

D. Channel Re-assignment Strategy

To adapt to the changing interference characteristics, the CAS periodically re-assigns channels. The periodicity depends ultimately on how frequently interference levels in the mesh network are expected to change. If a large number of interfering devices in the vicinity of the mesh network are expected to be short-lived, the invocation rate should be increased. On the other hand, if a majority of the interfering devices are likely to be long-lived, the invocation rate can be decreased.

References

1. Akyildiz, I.F., Wang, X., Wang, W.: Wireless mesh networks: a survey. *Comput. Netw. ISDN Syst.* 47(4) (2005)
2. Gupta, P., Kumar, P.R.: The Capacity of Wireless Networks. *IEEE Transactions on Information Theory* 46(2) (2000)
3. Kyasanur, P., Vaidya, N.H.: Capacity of Multi-Channel Wireless Networks: Impact of Number of Channels and Interfaces. In: *MOBICOM* (2005)
4. Gong, M.X., Midkiff, S.F., Mao, S.: A Combined Proactive Routing and MultiChannel MAC Protocol for Wireless Ad Hoc Networks. In: *Broadnets* (2005)
5. Subramanian, A.P., Gupta, H., Das, S.R., Cao, J.: "Minimum Interference Channel Assignment in Multiradio Wireless Mesh Networks. *IEEE Transactions on Mobile Computing* 7(12) (December 2008)

IEC 61850: Goose Messaging Implementation for MPR

Hemalata M. Shingate, Srija Unnikrishnan, and Sudarshan Rao

F.R.C.R.C.E. Bandra
Hema-shingate@powai.ltindia.com
F.R.C.R.C.E. Bandra, Mumbai
srija@frcrce.ac.in
L & T -Emsys, Mumbai
sr-emsys@powai.ltindia.com

Abstract. An electrical substation is a subsidiary station of an electricity generation, transmission and distribution system where voltage is transformed from high to low or the reverse using transformers. Substation communication plays a vital role in power system operation. The power industry is changing; there is more or less a global electrical power market with international power companies. Thus, national differences in for example communication protocols, special substation automation solutions etc, are an economic obstacle when the international electric companies are seeking even better productivity and economic performance. It is therefore necessary to design Substation Automation Systems that are more cost effective to operate and maintain and that ensure quicker returns on investment than in the past. It is therefore necessary to design Substation Automation Systems that are more cost effective to operate and maintain and that ensure quicker returns on investment than in the past. This new perspective with reference to the conceptual design and engineering of modern control systems, has been the driving force for the new standard IEC 61850, with a global approach on communication and information handling.

GOOSE is an acronym for Generic Object Orientated System-wide Events. It aims to replace the conventional hardwired logic necessary for intra-IED coordination with station bus communications. Upon detecting an event, the IED(s) use a multi-cast transmission to notify those devices that have registered to receive the data. The performance requirements are stringent – no more than 4ms is allowed to elapse from the time an event occurs to the time of message transmission. The main objective of the paper is to provide a framework for Substation Engineers to simulate GOOSE messages for fast transmission of substation events, such as commands, alarms, indications, as messages which take advantage of Ethernet and supports real-time behavior with Motor protection Relay using KALKI protocol Gateway Lite for Substation Automation project. To generate GOOSE messages with Embedded Artists LPC2468 OEM Board.

Keywords: IEC 61850, MPR, GOOSE messaging, communication, Substation.

1 Introduction

IEC 61850 is the new Ethernet-based international standard for communication in power generation facilities and substations. The goal of this standard is to integrate

all of the protection, control, measurement, and monitoring functions within a substation, and to provide the means for high-speed substation protection applications, interlocking and intertripping. These devices are generally referred to as Intelligent Electronic Devices (IED). [1] Over the last decade, the “digitization” of the electron enterprise has grown at exponential rates. Utility, industrial, commercial, and even residential consumers are transforming all aspects of their lives into the digital domain.[7] Moving forward, it is expected that every piece of equipment, every receptacle, every switch, and even every light bulb will possess some type of setting, monitoring and/or control. In order to be able to manage the large number of devices and to enable the various devices to communicate with one another, a new communication model was needed. That model has been developed and standardized as IEC61850 – Communication Networks and Systems in Substations¹. This paper looks at the needs of next generation communication systems and provides an overview of the IEC61850 protocol and how it meets these needs.

2 Scope and Outline of IEC 61850

The stated scope of IEC 61850 was communications within the substation. The document defines the various aspects of the substation communication network in 10 major sections as shown in Table 1 below.

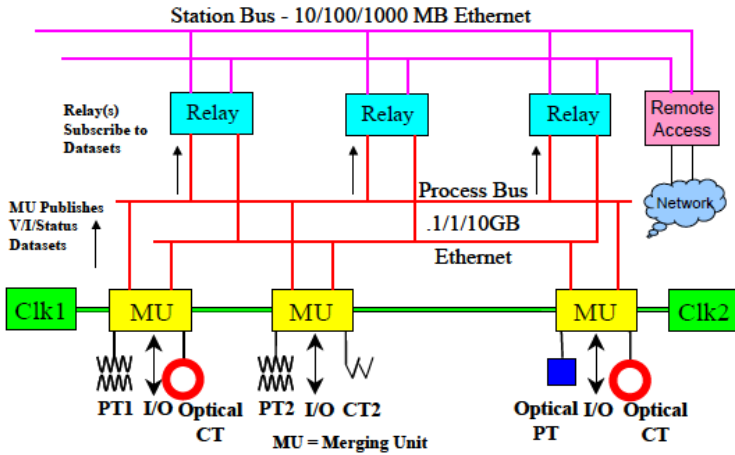
Table 1. Structure of the IEC 61850 standard

Table 1	
Part #	Title
1	Introduction and Overview
2	Glossary of terms
3	General Requirements
4	System and Project Management
5	Communication Requirements for Functions and Device Models
6	Configuration Description Language for Communication in Electrical Substations Related to IEDs
7	Basic Communication Structure for Substation and Feeder Equipment
7.1	- Principles and Models
7.2	- Abstract Communication Service Interface (ACSI)
7.3	- Common Data Classes (CDC)
7.4	- Compatible logical node classes and data classes
8	Specific Communication Service Mapping (SCSM)
8.1	- Mappings to MMS(ISO/IEC 9506 – Part 1 and Part 2) and to ISO/IEC 8802-3
9	Specific Communication Service Mapping (SCSM)
9.1	- Sampled Values over Serial Unidirectional Multidrop Point-to-Point Link
9.2	- Sampled Values over ISO/IEC 8802-3
10	Conformance Testing

3 IEC Substation Model

- 1) Function Hierarchy and Interfaces of IEC 61850: The three levels in the functional hierarchy are :

- i) **Process level:** This level includes switchyard equipments such as CTs / PTs, Remote I/O, actuators, etc.
- ii) **Bay level:** Bay level includes protection and control IEDs of different bays.
- iii) **Station level:** The functions requiring data from more than one bay are implemented at this level.



Clk1: Clock; PT: Potential Transformer; CT: Current Transformer

Fig. 1. IEC61850 Substation Model

4 The Approach of IEC 61850

OSI-7 Layer Communication System: The approach of IEC 61850 based on the separation of the object model with its data and services from the communication, i.e. from the ISO/OSI seven layer stack (Figure 2).

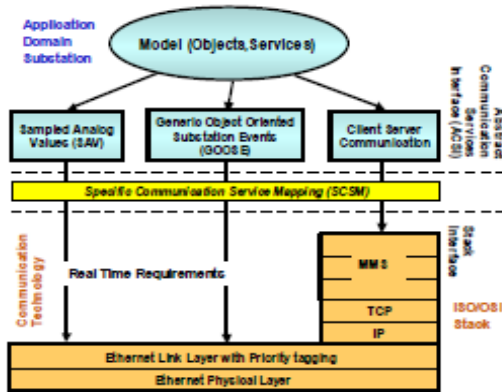


Fig. 2. Approach of IEC 61850: Split of data model and communication (7 layer ISO/OSI stack simplified, especially for MMS layers)

IEC 61850 uses OSI-7 layer stack for communication and divide it in three groups as shown in Fig.3. The seven types of messages are mapped into different communication stacks. The raw data samples (type 4) and GOOSE messages (type1) are time critical and are, therefore, directly mapped to low-level Ethernet layer. This gives the advantage of improved performance for real time messages by shortening the Ethernet frame (no upper layer protocol overhead) and reducing the processing time. The medium speed message (type 2), the command message with access control (type 7), the low speed message (type 3) and the file transfer functions (type 5) are mapped to MMS protocol which has a TCP/IP stack above the Ethernet layer. The time synchronization messages (type 6) are broadcasted to all IEDs in substation using UDP/IP.

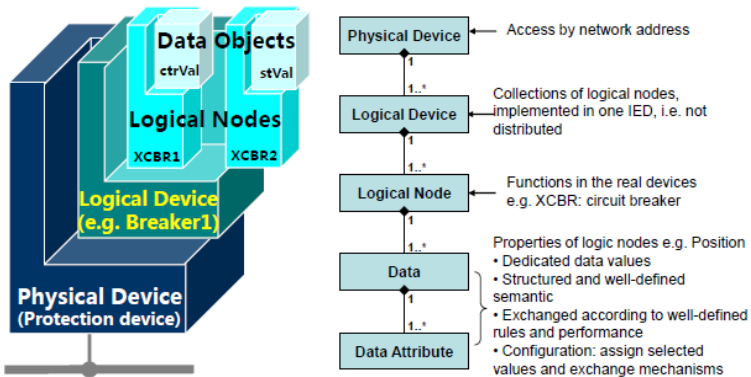


Fig. 3. IEC 61850 SCL - more than interoperable data exchange between engineering tools

Object Models

The IEC 61850 standard relies heavily on the Abstract Communication Service Interface (ACSI) model to define a set of service and the responses to those services. In terms of network behavior, abstract modeling enables all IEDs to act identically. These abstract models are used to create objects (data items) and services that exist independently of any underlying protocols. These objects are in conformance with the common data class (CDC) specification IEC 61850 7 3, which describes the type and structure of each element within a logical node. CDCs for status, measurements, controllable analogs and statuses, and settings all have unique CDC attributes. Each CDC attribute belongs to a set of functional constraints that groups the attributes into specific categories such as status (ST), description (DC), and substituted value (SV). Functional constraints, CDCs and CDC attributes are used as building blocks for defining Local Nodes.[2]

Data Mapping

Device data is mapped to IEC 61850 Logical Nodes (LN). In this project work we have mapped MODBUS formatted data into IEC 61850 Logical Nodes (LN).

File Services

Substation Configuration Language (SCL) is an XML-based configuration language used to support the exchange of database configuration data between different tools, which may come from different manufacturers.

There are four types of SCL files:

1. IED Capability Description file (.ICD)
2. System Specification Description (.SSD) file
3. Substation Configuration Description file (.SCD)
4. Configured IED Description file (.CID)

5 Description of the Application

GOOSE (Generic Object Oriented Substation Event):

The Generic Object Oriented Substation Event (GOOSE) object within IEC 61850 is for high speed control messaging. IEC 61850 GOOSE automatically broadcasts messages containing status, controls, and measured values onto the network for use by other devices. IEC 61850 GOOSE sends the messages several times, increasing the likelihood that other devices receive the messages. A GOOSE report enables high speed trip signals to be issued with a high probability of delivery.

Publisher-Subscriber Model:

To begin, the GOOSE message is not addressed by the sender to a particular receiving relay. Rather, it is sent as a broadcast (actually multicast) message that goes onto the LAN with identification of who the sender is, and with the identification of the specific message so that its point contents can be determined by listeners. There is no destination address. Every other relay and IED on the LAN can see the message, and decide on its own whether it needs to look at the contents of this message. [3][13]. The transmitting IED is called the publisher, and any other relay or IED that is configured to look for and use this particular message is called a subscriber. GOOSE messaging is also an unconfirmed service. Because of this, the publisher must keep on filling the LAN with updated GOOSE messages, and the burden of catching them falls to the individual subscribers.[6]

6 Project Development

Block Diagrams :

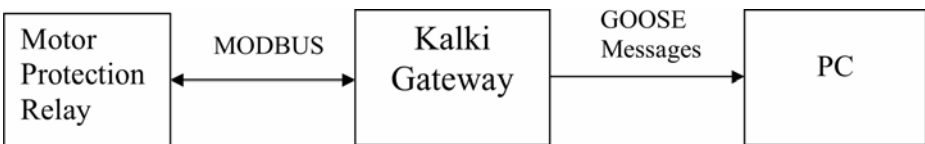


Fig. 4. Block diagram to generate GOOSE messages with Kalki Gateway

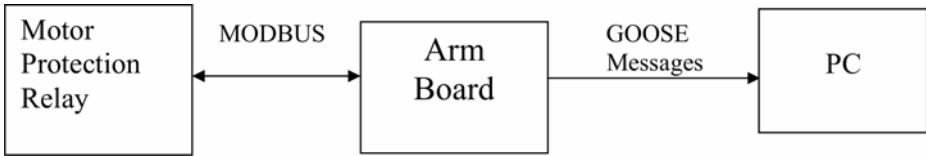


Fig. 5. Block diagram to generate GOOSE messages with ARM Board

7 Simulation and Results



Fig. 6. GOOSE demonstration setup using KALKI Protocol GatewayLite

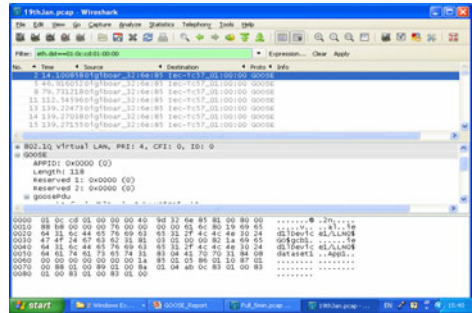


Fig. 7. GOOSE Messages observed with Wireshark [4]

Table 2. GOOSE packet Analysis with 1st Bit 0 to 1

Sr No.	Time(ms)	Time allowed to live(ms)	Sqnum
1	139	4	0
2	139	4	1
3	139	8	2
4	139	16	3
5	139	32	4
6	139	64	5
7	139	128	6
8	139	256	7
9	139	512	8
10	139	1024	9
11	140	2048	10
12	141	4096	11
13	143	8192	12
14	147	16384	13
15	156	32768	14
16	172	65536	15
17	205	65536	16
18	238	65536	17

8 Conclusion

This project work presents opportunities for enhancing power system automation through the use of real time Ethernet. It explains the latest international standards and mechanisms available to power system engineers. It should be noted that none of the applications presented in this paper may individually be able to justify the investment of equipping the substation with Ethernet network capability. However, once the Ethernet has been justified, a GOOSE message offers a preferred method for exchange of real-time status information among multiple protection devices. IEC GOOSE message can be used to simplify substation wiring, reduce installation cost, and enhance overall protection system performance. When properly applied, IEC GOOSE message offers a powerful new tool in the power system protection and automation engineer's toolbox. IEC61850 will become the protocol of choice as utilities migrate to network solutions for the substations and beyond.

GOOSE Messages with Kalki Tech Gateway

- In this paper **EasyConnect - The configuration/ Diagnostics utility** tools used to configure the Kalki Protocol Gateways. EasyConnect generates a configuration file in Extensible Markup Language (XML) format as the output.
- ICD Manager a software engineering platform is used to build different IED configurations, data models and GOOSE messages.
- IEDScout - a universal IEC 61850 client is used to connect to any IEC 61850 device (server-Kalki Tech Gateway) and provides many useful functions for viewing the data model, reading and writing data, reporting, along with GOOSE/GSSE publishing and subscribing.
- The MPR has MODBUS interface. MODBUS coil status we can be viewed with IEDScout.
- If coil status changes then Gateway sends GOOSE messages which we can observe with Wireshark - a Network Packet Analyser.
- On the occurrence of any change of state, an IED will multicast a high speed, Generic Object Oriented Substation Event (GOOSE) report by exception, typically containing the double command state of each of its status inputs, starters, output elements and relays, actual and virtual. This report is re-issued sequentially, typically after the first report, again at short interval. Initially 8-10 GOOSE messages arrives. Then next packets comes at 2ms, 4ms, 8ms & then GOOSE messages are repeated after every 33ms.

References

- [1] IEEE PSRC, WG H5, Application of peer-to-peer communications for protective relaying, web published report, <http://www.pes-psrc.org/h/H5doc.zip>
- [2] IEEE PSRC, WG H6, Application Considerations of UCA2 for Substation Ethernet LocalArea Network Communication for Protection and Control, web published report, <http://www.pes-psrc.org/h/>
- [3] IEC 61850-7-2, Communication networks and systems in substations – Part 7-2: Basic communication structure for substation and feeder equipment

- [4] <http://www.wireshark.org/>
- [5] <http://www.sisconet.com/>
- [6] <http://www.iec.ch/>
- [7] <http://library.abb.com/>

Performance Analysis of Energy Efficient Routing Algorithms for Adhoc Network

Dhiraj Nitnaware and Ajay Verma

Institute of Engineering & Technology, DAVV, Indore, India
dnitnaware.iet@rediffmail.com, ajayrt@rediffmail.com

Abstract. In the present paper, the performance of various energy efficient algorithms is being analyzed. In the previous papers, we have developed some energy efficient algorithm for MANET and analysis there performances. In [1] and [2], we have developed energy constraint gossip based protocol (ECG_AODV) and analyzed its energy performance as compared with AODV. Energy constraint Node cache based algorithm called ECNC_AODV were developed in [3]. In [4] and [5], we have proposed Energy based Gossip algorithm called EBG_AODV for MANET. In all the above work, we have analyzed the behavior of proposed algorithm and compare it with AODV under CBR as well Pareto traffic sources. With the help of simulation results, we have concluded that with above algorithm, there is an energy as well as overhead reduction up to 15%-35% without affecting the delivery ratio as compared to AODV. In this paper, we have simulated the above algorithm with new algorithm which is based on node cache plus gossip. The overall performance of EBG protocol is better as compared to other protocols in terms of energy consumption and routing overhead without affecting the delivery ratio as per simulation result.

Keywords: Energy efficient algorithms, energy consumption, routing overhead, delivery ratio, NS-2 simulator.

1 Introduction

MANET is a multi-hop, infrastructureless wireless network with limited battery power and bandwidth. Reactive protocols like AODV [11] and DSR [12] uses flooding technique to forward RREQ packets during route discovery process and RERR (route error) packets during route maintenance process. Due to flooding, many routing packets are unnecessarily increases which not only increases the energy consumption of each node but routing overhead as well.

In the previous work, Dhiraj et. al. [1] and [2] has proposed an energy constraint gossip based routing protocol in which the intermediate nodes forward the RREQ packets with some set probability k based on the energy status of the node. In [3], the author have proposed an energy efficient algorithm where the intermediate nodes will forward the RREQ packets based on the energy status and if the node is previously involved in data transfer i.e. if node cache.

In the paper [4] and [5], we have proposed another efficient algorithm in which the intermediate nodes will gossip with probability based on the current energy status, for example if remaining energy is 80% then gossip probability $k = 0.8$ and if 70% then

$k = 0.7$ and so on. This paper concentrates on the performance analysis of various energy efficient algorithms. The performance parameters taken for analysis are energy consumption due to routing packets, normalized routing overhead and delivery ratio (throughput).

The paper is organized as: Section 2 discuss the working of various routing protocol and proposed new protocols. The simulation model is presented in section 3. The simulation results are shown in section 4. Sections 5 describe conclusion and future scope.

2 Routing Protocol

This section gives the brief description on the working of various energy efficient algorithms and proposed algorithm.

2.1 The Ad Hoc on Demand Distance Vector

AODV protocol [11] is a reactive routing protocol which finds route to destination when demanded. It consists of routing table which contains sequence number and next hops information that helps to differentiate between stale and fresh routes. The connectivity between the nodes is maintaining using *Hello* messages. It has two processes: Route Discovery and Route Maintenance process. In route discovery process, the source node generate RREQ packet and floods it to the neighbouring nodes. The neighbouring nodes will flood to there neighbour and so on. When the packet reaches the destination node, it replies by generating RREP (Route Reply) packet. In route maintenance process, if the link is broken then the source node is being notified using RERR (Route Error) message.

2.2 Energy Constraint Gossip Based AODV (ECG_AODV)

In this algorithm [1] and [2], the intermediate nodes with energy higher than the set threshold are used in the route discovery process. These nodes will forward the RREQ packets with some associated probability k . If $E_n > E_{th}$, then n^{th} node forward RREQ packet with probability k provided that the numbers of neighbours are more than one, otherwise forward RREQ packet with probability 1 with one neighbour. If $E_n < E_{th}$, then the node will drop RREQ packet.

Here E_n is the current node energy, E_{th} is the set energy threshold and $k = 0.7$.

2.3 Energy Constraint Node Cache Based AODV (ECNC_AODV)

The main aim of this proposed algorithm is [3] to reduce the number of routing packets generated due to flooding method so that there should be reduction in energy consumption; routing overhead and increase in network lifetime could be achieved. The node which previously involved in data transfer is called as node caching. In this algorithm, the intermediate node will forward RREQ packets if it is node caching. The condition for this is as follows:

If $E_n > E_{th}$ and $T - \tau \leq T(N)$, then n^{th} node will forward RREQ packet.

If $E_n < E_{th}$, then drop RREQ packet.

where E_n is the current node energy, E_{th} is the set energy threshold (normal zone is considered), T is the current time, $T(N)$ is the time when last data packet transmitted through node N and τ is another small set time threshold which decides the memory of the node and is taken as 30sec.

2.4 Energy Based Gossip (EBG_AODV)

In this algorithm [4] and [5], during the route discovery process, the intermediate nodes forward the RREQ packets with probability k . This probability k is calculated on the basis of current energy status of that node. If the nodes remaining energy is 80% of the initial energy then $k = 0.8$, if 75% then $k = 0.75$, if 60% then $k = 0.6$ and so on.

2.5 Proposed Energy + Node Cache + Gossip (E+NC+G_AODV)

Here the intermediate nodes will forward RREQ packets if

$$E_n > E_{th} \quad \text{and} \quad T - \tau \leq T(N)$$

n^{th} node will forward RREQ packet with probability $k = 0.7$. If $E_n < E_{th}$, then the node will drop RREQ packet.

The intermediate node first checks its current energy. If it is more than the set threshold then the node will see the node caching. If it is previously involved in data transfer then it will gossip with $k = 0.7$. Here $\tau = 30\text{sec}$.

3 Simulation Model

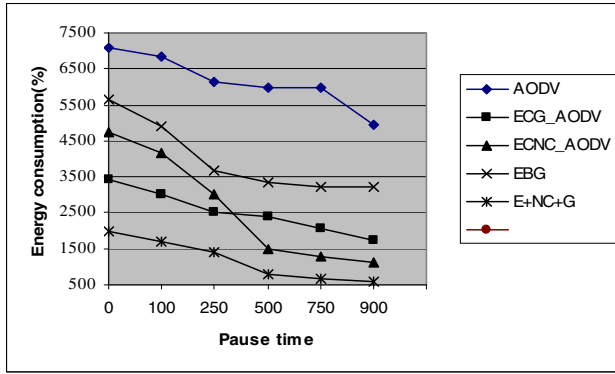
In simulation model, we have taken 50 nodes that are randomly distributed in a region of 1000m X 1000m with 30 number of connection. The energy model have NIC card consists of radio range of 250m, 2Mbps data rate with initial energy supplied to each node is 200J and power consumed during transmission and reception is 1.65W and 1.1W respectively. The traffic model used is CBR (Constant Bit Rate) with packet size of 512 bytes, sending rate 64 packets/s and simulation time of 900s. The simulation is done with the help of ns-2 [14] and traffic model is generated using cbreng.tcl [15].

4 Results

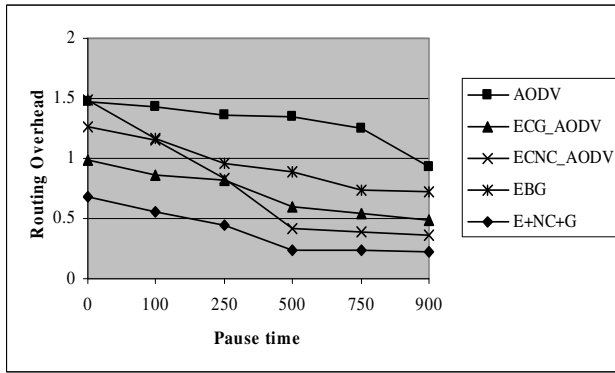
We have evaluated (i) Energy consumption due to routing packets (ii) Routing overhead and (iii) Delivery ratio for analysis between various protocols and following results was obtained.

4.1 By Varying Pause Time

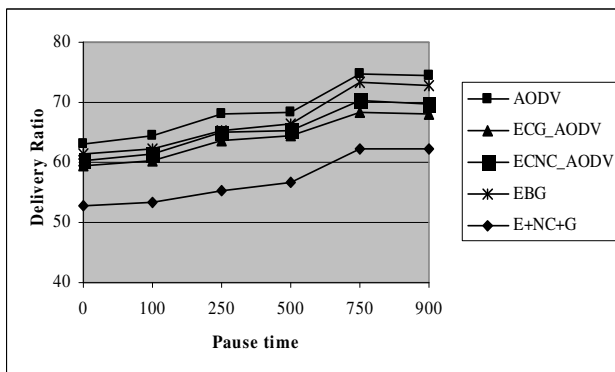
Figure 1a, 1b and 1c shows the total energy consumed (Joules) due to transmission and reception of control packets, routing overhead and delivery ratio by various protocols. From the result, we observed that energy consumption of E+NC+G is 35% less but delivery ratio is also 12% less as compared to AODV.



(a)



(b)



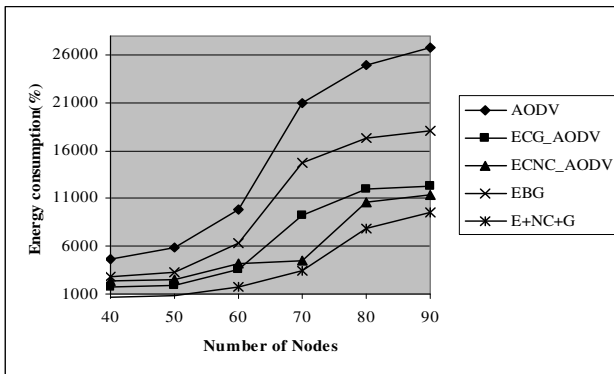
(c)

Fig. 1. (a) Energy consumption Versus Pause Time, (b) Routing Overhead versus Pause Time, (c) Delivery Ratio versus Pause Time

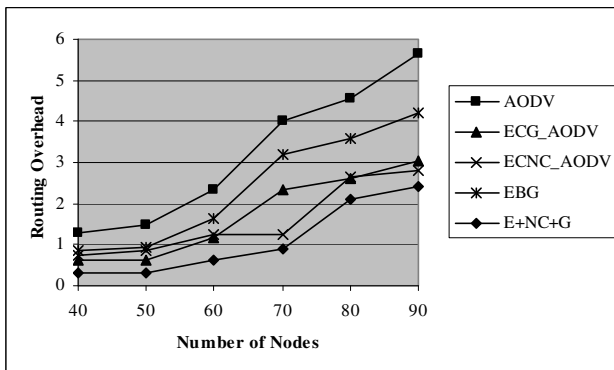
The best performance is shown by EBG protocol. There is 20% reduction in energy consumption and routing overhead with almost same delivery ratio as compared to AODV.

4.2 By Varying Number of Nodes

Figure 2a, 2b and 2c shows the total energy consumed (Joules) due to transmission and reception of control packets, routing overhead and delivery ratio by various protocols. E+NC+G shows 35% less energy and overhead reduction but delivery ratio is also 15% less as compared to AODV. Also EBG protocol shows 25% reduction in energy consumption, routing overhead and 2% less delivery ratio, ECG_AODV and ECNC_AODV shows 20% energy and overhead reduction with 5% less delivery ratio.

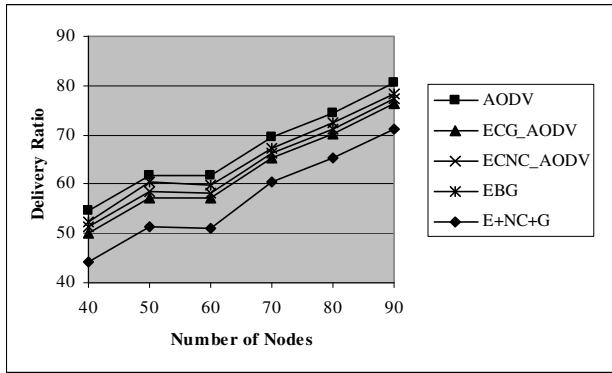


(a)



(b)

Fig. 2. (a) Energy consumption versus No. of Nodes, (b) Routing Overhead versus No. of Nodes, (c) Delivery Ratio versus No. of Nodes

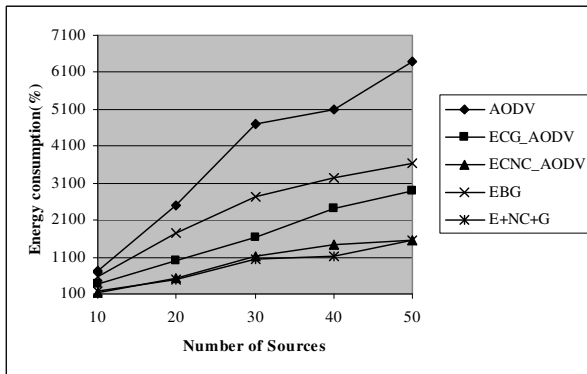


(c)

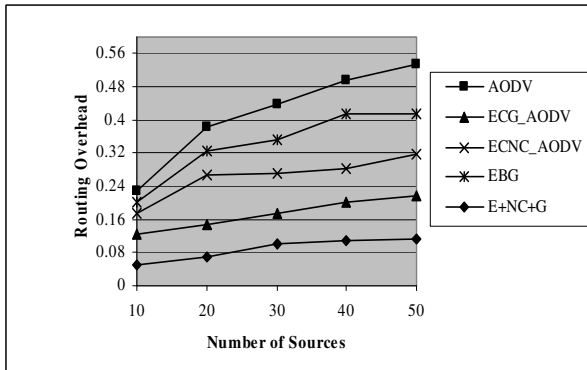
Fig. 2. (continued)

4.3 By Varying Number of Sources

In figure 3a, 3b and 3c, the performance of E+NC+G is best with 30% reduction in energy and overhead and with 10% more delivery ratio as compared to AODV.

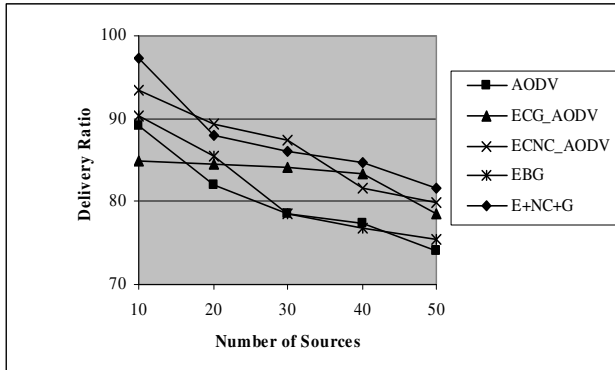


(a)



(b)

Fig. 3. (a) Energy consumption versus No. of Source, (b) Routing Overhead versus No. of Source (c) Delivery Ratio versus No. of Sources



(c)

Fig. 3. (continued)

ECNC_AODV also shows 25% reduction in energy and overhead with 5% more delivery ratio. Also EBG protocol shows 10% reduction in energy consumption, routing overhead and 2% more delivery ratio.

5 Conclusion and Future Scope

With the help of simulation results, table 1, 2 and 3 were obtained. Positive sign shows the better performance and negative sign shows poor performance as compared to AODV protocol.

Thus we conclude that:

(i) If 1%-2% delivery ratio is tolerable then EBG protocol shows the better performance in high mobility, scalability and network size (pause time, speed, number of nodes and grid area).

(ii) If 5%-8% delivery ratio is tolerable then ECNC_AODV and ECG_AODV protocol as a function of pause time, speed, number of nodes and grid area.

(iii) The performance of E+NC+G is better among other protocols if there is change in number of sources and sending rate while for other parameters the delivery is with 10%-15% less.

Table 1. Energy Consumption (%)

Protocols	Pause Time	Speed	No. of Nodes	No. of Sources	Grid Area	Sending Rate
E+NC+G	+35%	+35%	+35%	+30%	+35%	+30%
ECNC_AODV	+30%	+30%	+20%	+25%	+25%	+25%
ECG_AODV	+25%	+25%	+20%	+25%	+25%	+25%
EBG	+20%	+20%	+25%	+10%	+25%	+10%

Table 2. Routing Overhead (%)

Protocols	Pause Time	Speed	No. of Nodes	No. of Sources	Grid Area	Sending Rate
E+NC+G	+35%	+35%	+35%	+30%	+35%	+30%
ECNC_AODV	+30%	+30%	+20%	+25%	+25%	+25%
ECG_AODV	+25%	+25%	+20%	+25%	+25%	+25%
EBG	+20%	+20%	+25%	+10%	+25%	+10%

Table 3. Delivery Ratio (%)

Protocols	Pause Time	Speed	No. of Nodes	No. of Sources	Grid Area	Sending Rate
E+NC+G	-12%	-12%	-15%	+10%	-15%	+10%
ECNC_AODV	-8%	-8%	-5%	+5%	-5%	+5%
ECG_AODV	-5%	-5%	-5%	+5%	-5%	+5%
EBG	-1%	-1%	-2%	-2%	-2%	-2%

The overall performance of EBG protocol is better as compared to other protocols in terms of energy consumption and routing overhead without affecting the delivery ratio. The delivery ratio of all the above protocols can be increased by proper selection of parameter such as τ , *energy threshold* and *gossip probability* k .

In future, we want to increase the delivery ratio and also to simulate the above protocols under real traffic like Pareto traffic and compare it with CBR traffic.

References

1. Nitnaware, D., Karma, P., Verma, A.: Energy Constraint Gossip Based Routing Protocol for MANETs. In: Proc. of IEEE International Conference on Advances in Computer Vision and Information Technology (ACVIT 2009), Aurangabad, December 16-19, pp. 423–430 (2009)
2. Nitnaware, D., Karma, P., Verma, A.: Performance Analysis of Energy Constraint Gossip Based Routing Protocol under Stochastic Traffic. In: Proc. of IEEE International Conference on Emerging Trends in Engineering & Technology (ICETET 2009), Nagpur, December 16-18, pp. 1110–1114 (2009)
3. Nitnaware, D., Verma, A.: Energy Constraint Node Cache Based Routing Protocol for Adhoc Network. International Journal of Wireless & Mobile Networks (IJWMN) 22, 77–86 (2010) ISSN: 0975-3834 (Online); 0975-4679 (Print)
4. Nitnaware, D., Verma, A.: Energy Based Gossip Routing Algorithm for MANETs. In: Proc. of ACEEE International Conference on Information, Telecommunication and Computing (ITC 2010), Cochin, Kerala, March 12-13, pp. 23–27 (2010)
5. Nitnaware, D., Karma, P., Verma, A.: Performance Evaluation of Energy Based Gossip Routing Algorithm for On-OFF Source Traffic. In: Proc. of Springer International Conference on Contours of Computing Technology (THINKQUEST 2010), Mumbai, March 13-14, pp. 327–331 (2010)

6. Nitnaware, D., Verma, A.: Performance Evaluation of Energy Consumption of Reactive Protocols under Self- Similar Traffic. *International Journal of Computer Science & Communication (IJCS)* 1(1) (February 2010) ISSN: 0973-4414
7. Nitnaware, D., Verma, A.: Energy Evaluation of Proactive and Reactive Protocol for MANET Under ON/OFF Source Traffic. In: *Proceeding of ACM International Conference on Advances in Computing, Communication and Control (ICAC3 2009)*, Mumbai, January 23-24, pp. 451–455 (2009)
8. Nitnaware, D., Verma, A.: Energy Evaluation of Two Reactive Protocols under ON/OFF Source Traffic. In: *Proceeding of International Conference on Advance Computer Technologies (ICACT 2008)*, GRIET, Hyderabad, December 26-27, pp. 745–749 (2008)
9. Haas, Z.J., Halpern, J.Y.: Gossip Based Ad Hoc Routing. *IEEE Transactions on Networking* 14(3) (June 2006)
10. Frikha, M., Ghandour, F.: Implementation and Performance Evaluation of an Energy Constraint Routing Protocol for MANET. In: *3rd International Conference of Telecommunications (AICT 2007)*. IEEE, Los Alamitos (2007)
11. Perkins, C.E., Royer, E.M., Das, S.: Ad-Hoc on Demand Distance Vector Routing (AODV), draft-ietfmanet-aodv-05.txt (March 2000)
12. Johnson, D.B., Maltz, D.A., Hu, Y.C.: DSR for Mobile Ad Hoc Network. Internet Draft, draft-ietfmanet-drs-09.txt (July 2003)
13. Mahesh, N., Sundararajan, T.V.P., Shanmugam, A.: Improving performance of AODV Protocol using Gossip based approach. In: *IEEE International Conference on Computational Intelligence and Multimedia Applications (2007)*
14. Network Simulator, ns-2, <http://www.isi.edu/nsnam/-ns/>
15. <http://www.isi.edu/nsnam/ns/tutorial/>
16. Cano, J.C., Manzoni, P.: A Performance Comparison of Energy Consumption for Mobile Ad Hoc Network Routing Protocols. In: *Proceeding of 8th International Symposium on Modeling, Analysis and Simulation of Computer & Telecommunication System (2000)*
17. Johnson, D.B., Maltz, D.A., Broch, J.: DSR for Mobile Ad Hoc Network (November 1999)
18. Ehsan, H., Uzmi, Z.A.: Performance Comparison of Ad Hoc Wireless Network Routing Protocols. In: *Proceeding of INMIC 2004, 8th International Multitopic Conference*, pp. 457–465. IEEE, Los Alamitos (2004)
19. Perkins, C.E., Bhagwat, P.: Highly Dynamic Destination Sequenced Distance Vector Routing (DSDV) for Mobile Computers. In: *Proceeding of ACM SIGCOMM (October 1994)*

ECG Signal Compression Using Different Techniques

K. Ranjeet, A. Kumar, and R.K. Pandey

Indian Institute of Information Technology Design and Manufacturing,
Jabalpur-482005, MP (India)
ranjeet281@gmail.com, anilkdee@gmail.com,
wavelet_r@yahoo.co.in

Abstract. In this paper, a transform based methodology is presented for compression of electrocardiogram (ECG) signal. The methodology employs different transforms such as Discrete Wavelet Transform (DWT), Fast Fourier Transform (FFT) and Discrete Cosine Transform (DCT). A comparative study of performance of different transforms for ECG signal is made in terms of Compression ratio (*CR*), Percent root mean square difference (*PRD*), Mean square error (*MSE*), Maximum error (*ME*) and Signal-to-noise ratio (*SNR*). The simulation results included illustrate the effectiveness of these transforms in biomedical signal processing. When compared, Discrete Cosine Transform and Fast Fourier Transform give better compression ratio, while Discrete Wavelet Transform yields good fidelity parameters with comparable compression ratio.

Keywords: ECG, Compression, DWT, DCT, Huffman encoding, FFT.

1 Introduction

An electrocardiogram (ECG) is the graphical representation of electrical impulses due to ionic activity in the cardiac muscles of human heart. It is an important physiological signal which is exploited to diagnose heart diseases because every arrhythmia in ECG signals can be relevant to a heart disease [1, 2]. ECG signals are recorded from patients for both monitoring and diagnostic purposes. Therefore, storage of computerized is become necessary. However, the storage has limitation which made ECG data compression as an important issue of research in biomedical signal processing. In addition to these, there are many advantages of ECG compression such as transmission speed of real-time ECG signal is enhanced and is also economical.

An ECG signal contains steep slopes QRS complexes and smoother P and T waves. It is recorded by applying electrodes to various locations on the body surface and connecting them to a recording apparatus. There are certain amounts of sample points in ECG signal which are redundant and replaceable. ECG data compression is achieved by elimination of such redundant data sample points. Therefore, in early stage of research, several methods for ECG compression were introduced to achieve good compression ratio with preserving the relevant signal information. These algorithms were classified into three categories [3]: dedicated techniques such as AZTEC, FAN, CORTES, and turning point. These techniques were based on the detection and elimination of redundancies on direct analysis of the original signal, and

gives minimum distortion. In second category, all transform based techniques comes and compression is achieved based on spectral and energy distribution of the signal. Other hand, last technique is based on feature and parameter extraction in which some parameter such as measurement of probability distribution of original signal is extracted. During last the two decades, several efficient methods [4-9] have reported in literature, which involve compression of ECG signal without losing and preserving the relevant clinical information for the accurate detection and classification.

In past, researches have proposed many transform methods such as Discrete Cosine Transform (DCT), Fast Fourier Transform (FFT) and Discrete Wavelet Transform (DWT), due to that, there is drastically changed in the field of data compression. FFT is a discrete Fourier transform (DFT) algorithm which reduces the number of computations needed for N points from $2N^2$ to $2N \log_2 N$ [10]. DFT is used in Fourier analysis of a signal in frequency domain. Based on FFT, many methods [10, 11] have been proposed for analyzing and compressing ECG signal. The discrete cosine transform is widely exploited for data compression such as speech compression, image compression and ECG compression. DCT is calculated using the FFT algorithm as it is DFT. However, DCT gives the more weight to low-pass coefficients to high-pass coefficients. DCT gives nearly optimal performance in the typical signal having high correlations in adjacent samples. Several researchers [12-16] have developed unique algorithms for compression of ECG signal based on Discrete Cosine Transform. The detailed discussion on FFT and DCT for ECG compression is given in [10-16] and the references there in.

During the last decade, the Wavelet Transform, more particularly Discrete Wavelet Transform has emerged as powerful and robust tool for analyzing and extracting information from non-stationary signal such as speech signal and ECG signal due to the time varying nature of these signals. Non-stationary signals are characterized by numerous abrupt changes, transitory drifts, and trends. Wavelet has localization feature along with its time-frequency resolution properties which makes it suitable for analyzing non-stationary signals such as speech and electrocardiogram (ECG) signals. Recently, several other methods [17-25] have been developed based on wavelet or wavelet packets for compressing ECG signal.

In above context, therefore, this paper presents some new results based on transform technique such as DWT, FFT and DCT for ECG signal compression. The paper is organized as follows. A brief introduction has been provided in this section on the existing compression techniques of ECG signal. Section 2 discusses overview of different transforms such DWT, FFT and DCT. Section 3 presents the methodology of ECG compression based on these transforms. Finally, a comparison of results obtained with these transforms is carried out in Section 4, followed by concluding remarks in Section 5.

2 Techniques for ECG Signal Compression

In this paper, three transforms such FFT, DCT and DWT are employed for the ECG signal compression.

2.1 Fast Fourier Transform

A signal having periodic function of time can be analyzed or synthesized as a number of harmonically related to sine and cosine signals [10, 11]. A periodic signal $f(t)$ with period T_0 can be represented by Fourier series as

$$f(t) = A_0 + \sum_{n=1}^{\infty} a_n \cos(2\pi n t / T_0) + \sum_{n=1}^{\infty} b_n \sin(2\pi n t / T_0) \tag{1}$$

where, A_0 is the average, or mean value of signal a_n and b_n are the Fourier series coefficients. t is the time and n is the coefficient index. The above Fourier series coefficients are found by FFT.

$$f(t) = A_0 + \frac{1}{2} \sum_{n=1}^{\infty} (a_n - j b_n) e^{j 2\pi n t / T_0} \tag{2}$$

$$= \sum_{n=1}^{\infty} \alpha_n e^{j 2\pi n t / T_0} \tag{3}$$

where, α_n are complex coefficients. It is also expressed as

$$\alpha_n = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} f(t) e^{-j 2\pi n t / T_0} dt \quad n=0, \mp 1, \mp 2, \dots \tag{4}$$

For the sampled periodic signal, the discrete-time complex coefficients of the series are:

$$\alpha_n = \frac{1}{N} \sum_k^{N-1} f(k) e^{j 2\pi k n / N} \tag{5}$$

and

$$f(k) = \sum_k^{N-1} \alpha_n e^{-j 2\pi k n / N} \tag{6}$$

where, k is the discrete time index. N is the number of ECG signal samples. From equations (1), (2) and (4) give the Fourier series coefficients of Eq. (5) calculated using FFT technique. Since the ECG signal decomposition is assumed to be time varying due to cardiac disorders, Eq. (5) must be performed on each detected cycle. Fourier series coefficients used to synthesize the original signal is computed using Eq. (6) [12].

2.2 Discrete Cosine Transform

DCT has widely used for the data compression. In the signal decomposition based on DCT algorithms has four essential steps: dividing a signal in N sub-parts; DCT computation for each block; Thresholding & Quantization of the DCT coefficients; and encoding of the quantized DCT coefficients.

Discrete cosine transform is defined as

$$X(n) = \left(\frac{1}{N}\right)^{\frac{1}{2}} \sum_{i=0}^{N-1} x(i) \cos\left[\frac{\pi n}{2N}(2j+1)\right] \tag{7}$$

While the inverse IDCT is defined as:

$$x(i) = \left(\frac{1}{N}\right)^{\frac{1}{2}} \sum_{n=0}^{N-1} X(n) \cos\left[\frac{\pi i}{1N}(2n+1)\right] \tag{8}$$

DCT gives the decomposed coefficient of the original signal and it gives the more weight to low-pass coefficients to high-pass coefficients [12-16].

2.3 Discrete Wavelet Transform

Wavelets transform is a method to analyze a signal in time and frequency domain. DWT gives the multiresolution decomposition of a signal. There is three basic concept of multiresolution: subband coding, vector space and pyramid structure coding [18]. DWT decompose a signal at several n levels in different frequency bands. Each level decomposes a signal into approximation coefficients (low frequency band of processing signal) and detailed coefficients (high frequency band of processing signal) [19-25] as shown in Fig. 1

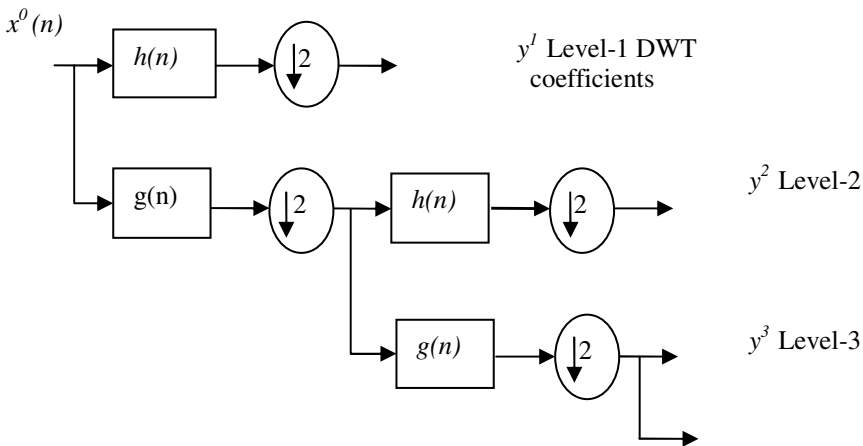


Fig. 1. Filter bank representation of DWT decomposition

At each step of DWT decomposition, there are two outputs: scaling coefficients $x^{j+1}(n)$ and the wavelet coefficients $y^{j+1}(n)$. These coefficients are given as

$$x^{j+1}(n) = \sum_{i=1}^{2n} h(2n-i)x^j(n) \tag{9}$$

and

$$y^{j+1}(n) = \sum_{i=1}^{2n} g(2n-i)x^j(n) \tag{10}$$

where, the original signal is represented by $x^0(n)$ and j show the scaling number. Here $g(n)$ and $h(n)$ represent the low pass and high pass filters, respectively. The output of scaling function is input of next level of decomposition, known as approximation coefficients. The approximation coefficients are low-pass filter coefficients and high-pass filter coefficients are detail coefficients of any decomposed signal.

3 Methodology for ECG Signal Compression

In this paper, the ECG compression is achieved using different transformation techniques such as FFT, DCT and DWT. The methodology of ECG compression based on transform is shown in Fig.2.

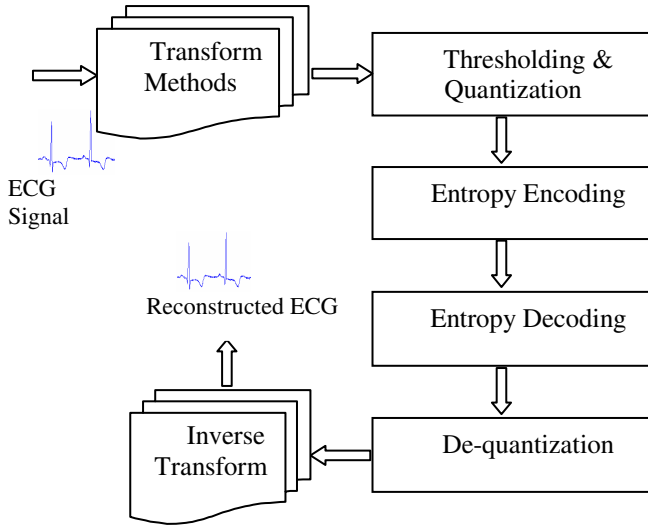


Fig. 2. Methodology for ECG signal compression

The algorithm of ECG compression is performing in three stages: (i) Transform calculation, (ii) Thresholding & Quantization, (iii) Entropy encoding. First, the signal (ECG signal) transformation is done with different transforms. Then, apply a threshold condition on the transform coefficients on the basis of energy packing efficiency of coefficients, which make a fixed percentage of coefficients equal to zero. Here, Global thresholding is used in which the threshold value is set manually, this value is chosen from transform analysis coefficient $(0 \dots x_{max}^j)$, where x_{max}^j is maximum coefficient. A detailed discussion on thresholding is given in [22, 24].

Further, uniform quantizer is employed on these coefficients. In quantization process, wavelet coefficients are quantized using uniform step size. The computation of step size depends on three parameters [13, 14]: maximum (M_{\max}) and minimum (M_{\min}) values in the signal matrix, and number of quantization level (L). Once these parameters are found, then step size (Δ)

$$\Delta = (M_{\max} - M_{\min}) / L \quad (11)$$

Then the input is divided into $L+1$ level with equal interval size ranging from M_{\min} to M_{\max} to plot quantization table. When quantization is done, then quantized values are fed to the next stage of compression and these three parameters defined above are stored in a file as they are required for creating the quantization table during reconstruction step. Detailed discussion on quantization is available in the references [6, 22, 24]. The actual compression is achieved at this stage and it is further achieved with the help of entropy encoding technique (Huffman). The quantized data contains same redundant data, means repeated data presents, it is waste of space. The way to overcoming this problem Huffman encoding is used. In this, the probabilities of occurrence of the symbols in the signal are computed. After that, these are arranged according to the probabilities of occurrence in descending order and build a binary tree and codeword table [24, 26, and 27]. Finally, compressed ECG signal is obtained.

4 Results and Discussion

In this paper, ECG signal compression is achieved using a methodology based on different transforms such as FFT, DCT and DWT. The performance of methodology algorithm can be evaluated by considering the fidelity of the reconstructed signal to the original signal. For this purpose, following fidelity assessment parameters [22-24] are considered:

Compression ratio (CR):

$$CR = \frac{\text{Number of significant wavelet coefficients}}{\text{Total number of wavelet coefficients}} \quad (12)$$

Percent root mean square difference (PRD):

$$PRD = \left(\frac{\text{Reconstructed noise energy}}{\text{Original signal energy}} \right)^{1/2} \times 100 \% \quad (13)$$

Mean square error (MSE):

$$MSE = \frac{1}{2} \sum_n |x(n) - y(n)|^2 \quad (14)$$

Maximum error (ME):

$$ME = \max_n |x(n) - y(n)| \quad (15)$$

Signal to noise ratio (SNR):

$$\begin{aligned}
 SNR &= 10 \log_{10} \left(\frac{\text{energy of input signal}}{\text{energy of the reconstructed error}} \right) \\
 &= 10 \log_{10} \left\{ \frac{\sum x^2(n)}{\sum |x(n) - y(n)|^2} \right\}
 \end{aligned} \tag{16}$$

ECG records have been obtained from MIT-BIH Arrhythmia Database (Physionet Bank) [28]. Here, FFT, DCT and DWT are employed for same ECG signal (MIT BIH ECG record 100) and the simulation results obtained in each case are included in Table 1. In all three cases, global thresholding is applied. In case DWT, different wavelet filters such as Haar, db7, db10, bior3.5, coif3, coif4, and coif 5 are used for ECG compression. Fig. 3 shows the plot of the original ECG signals (MIT-BIH record 100) and its reconstructed version. While in Fig. 4, a comparative analysis of different transforms is depicted.

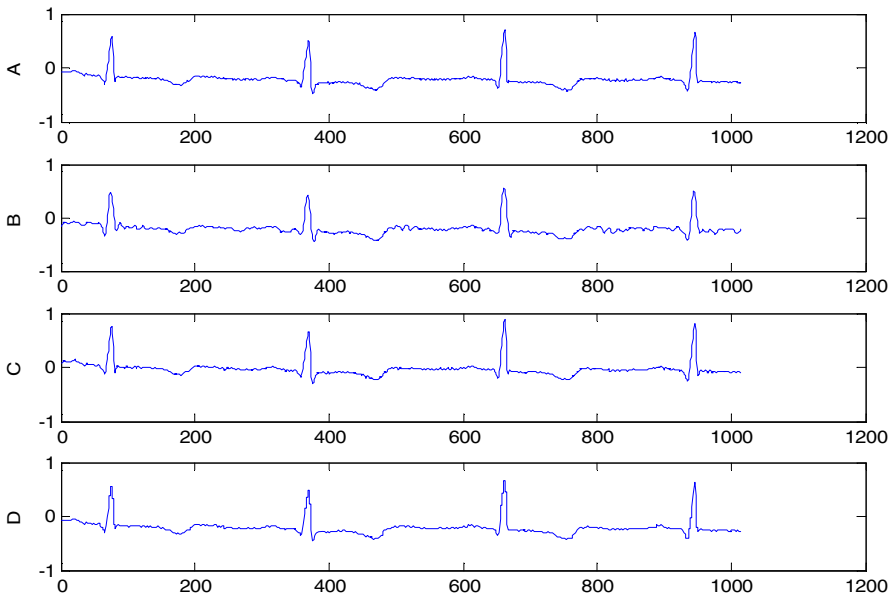


Fig. 3. (a) Original ECG signal, (b) Reconstructed signal using FFT, (c) Reconstructed signal using DCT, (d) Reconstructed signal using DWT

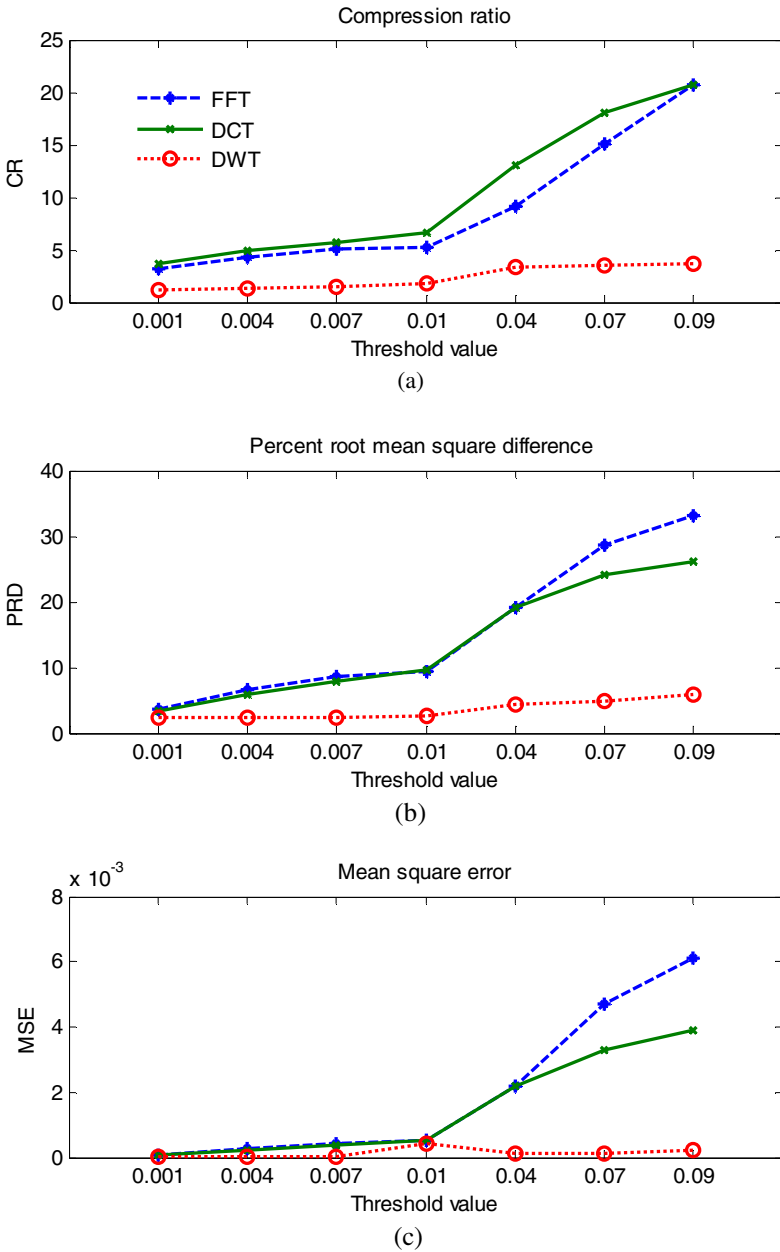
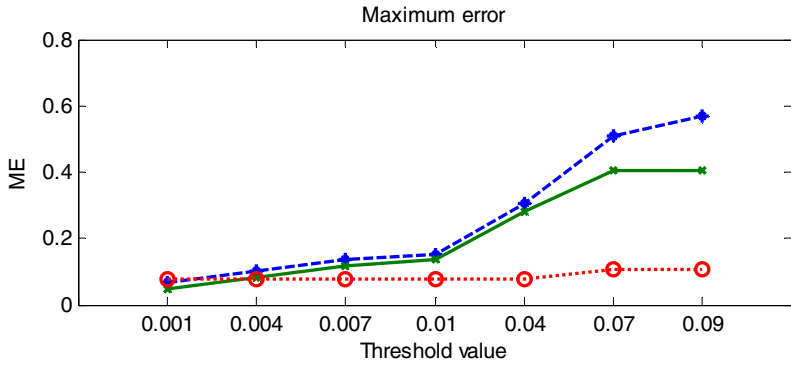
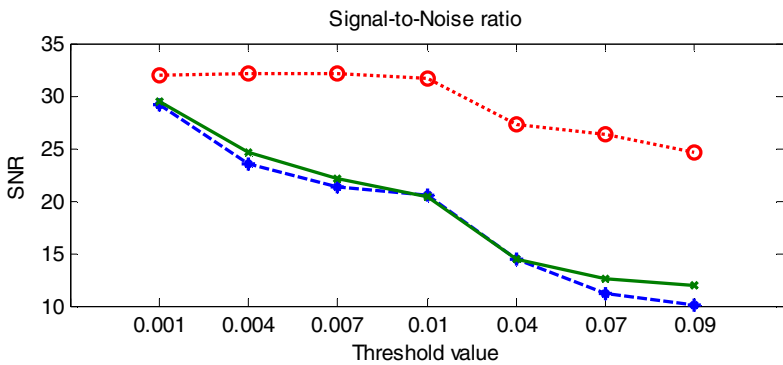


Fig. 4. A comparative analysis of the performance in different transforms(DCT, DWT and FFT) (a) Compression ratio (CR) (b) *PRD* (c) *MSE* (d) Maximum error (*ME*) (e) *SNR*



(d)



(e)

Fig. 4. (continued)

Table 1. Fidelity assessment parameters in different transforms

Type of Transform	CR	PRD	MSE	ME	SNR
FFT	5.31	9.34	5.32×10^{-4}	0.1537	20.62
DCT	6.58	9.53	5.45×10^{-4}	0.1353	20.45
DWT (Haar)	3.98	13.28	1.10×10^{-3}	0.2661	17.58
DWT (dB7)	3.86	9.55	5.47×10^{-4}	0.2269	20.43
DWT (dB10)	3.79	9.66	5.59×10^{-4}	0.2411	20.33
DWT (Bior 3.5)	3.88	11.68	8.05×10^{-4}	0.2660	18.75
DWT (Coif3)	3.81	8.46	4.34×10^{-4}	0.1788	21.43
DWT (Coif4)	3.76	8.28	4.15×10^{-4}	0.1714	21.62
DWT (Coif5)	3.68	8.17	4.062×10^{-4}	0.1680	21.72

It is evident from Table I and Fig. 4 that the good compression ratio can be obtained with these transforms with good fidelity measuring parameters. When compared, Discrete Cosine Transform and Fast Fourier Transform give better compression ratio, while Discrete Wavelet Transform yields good fidelity parameters with comparable compression ratio. The average compression ratio obtained in case of FFT and DCT are 5.31 and 6.58, respectively. While in case different wavelet filters, the compression ratio are 3.98, 3.86, 3.79, 3.88, 3.81, 3.76 and 3.68. Therefore, these transforms can be effectively used for ECG signal compression while preserving necessary clinical information.

5 Conclusions

In this paper, a transform based methodology is presented for ECG signal compression. A comparative study of performance of different transforms such as DCT, FFT and DWT for ECG compression is made. DWT decomposition is perfect to preserve clinical information, while DCT and FFT gives the high compression ratio. It is evident from the simulation results that these transforms can be effectively used for compression and analysis of ECG signal.

References

1. Afonso, V.X., Tompkins, W.J., Nguyen, T.Q., Luo, S.: ECG beat detection using filter banks. *IEEE Trans. Biomed. Eng.* 46, 192–202 (1999)
2. Afonso, V.X., Tompkins, W.J., Nguyen, T.Q., Michler, K., Luo, S.: Comparing stress ECG enhancement algorithms: with an introduction to a filter bank based approach. *IEEE Eng. Med. Biol. Mag.* 15, 37–44 (1996)
3. Ole-Aase, S., Nygaard, R., Husoy, J.H.: A comparative study of some novel ECG data compression technique. In: *NORSIG 1998*, pp. 273–276 (1998)
4. Jalaeddine, S.M.S., Hutchens, C.G., Strattan, R.D., Coberly, W.A.: ECG Data Compression Techniques-A Unified Approach. *IEEE Transactions on Biomedical Engineering* 37, 329–342 (1990)
5. Koski, A., Juhola, M.: Segmentation of digital signals on estimated compression ratio. *IEEE Transactions on Biomedical Engineering* 43, 928–938 (1996)
6. Mammen, C.P., Ramamurthi, B.: Vector quantization for compression of multichannel ECG. *IEEE Transactions on Biomedical Engineering* 37, 821–825 (1990)
7. Horspool, R. N., Windels, W. J.: An LZ approach to ECG compression, proceeding IEEE Symp. Computer-based Medical System, 71-76, 1994.
8. Aydin, M.C.: ECG data compression by sub-band coding. *IEEE Electronics Letters* 27, 359–360 (1991)
9. Wang, C.H., Liu, J., Sun, J.: Compression algorithm for electrocardiograms based on sparse decomposition. *Front. Electr. Electron. Eng. China* 4, 10–14 (2009)
10. Al-Nashash, H.A.M.: A dynamic Fourier series for the compression of ECG using FFT and adaptive coefficient estimation. *Med. Eng. Physics* 17, 197–203 (1995)
11. Weisstein, E.W.: Fast Fourier Transform, From MathWorld—A Wolfram Web Resource, <http://mathworld.wolfram.com/FastFourierTransform.html>

12. Al-Hinai, N., Neville, K., Sadik, A.Z., Hussain, Z.M.: Compressed Image Transmission over FFT-OFDM: A Comparative Study. In: Australasian Telecommunication Networks and Applications Conference, pp. 465–469 (2007)
13. Batista, L.V., Melcher, E.U.M., Carvalho, L.C.: Compression of ECG signals by optimized quantization of discrete cosine transform coefficients. *Medical Engineering & Physics* 23, 127–134 (2001)
14. Allen, V.A., Belina, J.: ECG data compression using the discrete cosine transform (DCT). *IEEE Proceedings, Computers in Cardiology*, 687–690 (1992)
15. Birney, K.A., Fischer, T.R.: On the Model Modeling of DCT and Subband Image for Data Compression. *IEEE Transactions on Image Processing* 4, 186–193 (1995)
16. Aggarwal, V., Patterh, M.S.: Quality Controlled ECG Compression using Discrete Cosine Transform (DCT) and Laplacian Pyramid (LP). In: *Multimedia, Signal Processing and Communication Technology, IMPACT 2009*, pp. 12–15 (2009)
17. Rajoub, B.A.: An Efficient Coding Algorithm for the Compression of ECG Signals Using the Wavelet Transform. *IEEE Transactions on Biomedical Engineering* 49, 355–362 (2002)
18. Mallat, S.G.: A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 11 (July 1989)
19. Kim, B.S., Yoo, S.K., Lee, M.H.: Wavelet-Based Low-Delay ECG Compression Algorithm for Continuous ECG Transmission. *IEEE Transaction on Information Technology in Biomedicine* 10, 77–83 (2006)
20. Ahmeda, S.M., Abo-Zahhad, M.: A new hybrid algorithm for ECG signal compression based on the wavelet transformation of the linearly predicted error. *Medical Engineering & Physics* 23, 117–126 (2001)
21. Kumari, R.S.S., Sadasivam, V.: A novel algorithm for wavelet based ECG signal coding. *Computers and Electrical Engineering* 33, 186–194 (2007)
22. Chen, J., Wang, F., Zhang, Y., Shi, X.: ECG compression using uniform scalar dead-zone quantization and conditional entropy coding. *Medical Engineering & Physics* 30, 523–530 (2008)
23. Blanco-Velasco, M., Cruz-Roldan, F., Godino-Llorente, J.I., Blanco-Velasca, J., Armien-Aparicio, C., Lopez-ferreras, F.: On the use of PRD and CR parameters for ECG compression. *Medical Engineering & Physics* 27, 798–802 (2005)
24. Manikandan, M.S., Dandapat, S.: Wavelet threshold based ECG compression using USZZQ and Huffman coding of DSM. *Biomedical Signal Processing and Control* 1, 261–270 (2006)
25. Brechet, L., Lucas, M.F., Doncarli, C., Farina, D.: Compression of Biomedical Signals With Mother Wavelet Optimization and Best-Basis Wavelet Packet Selection. *IEEE Transactions on Biomedical Engineering* 54, 2186–2192 (2007)
26. Huffman, D.A.: A method for the construction of minimum-redundancy codes. *Proceedings IRE* 40, 1098–1101 (1952)
27. Tanaka, H.: Data structure of Huffman codes and its application to efficient encoding and decoding. *IEEE Transactions Information Theory* 33, 154–156 (1987)
28. MIT-BIH arrhythmia ECG signal database

Efficient Content Based Video Retrieval for Animal Videos

Vijay Katkar¹ and Amit Barve²

¹ MPSTME, NMIMS University, Mumbai

² Fr. Conceicao Rodrigues College of Engineering, Mumbai

katkarvijayd@gmail.com, barve.amit@gmail.com

Abstract. Researchers are working on Content Based Video Retrieval (CBVR) for many years. But the available techniques for CBVR are not performing well for CBVR of Animal Videos. Thus in this paper we have proposed a new technique for CBVR of Animal Videos and experimental results are also provided with this paper to support it.

1 Introduction

With recent advances in multimedia technologies, processing speed and storage capacity more and more data are being captured, produced and stored. But if efficient techniques are not available to access the data then this data is hardly useful. Thus the research on efficient retrieval of video data is a very hot field.

Content-based Video Retrieval systems appear like a natural extension of Content-based Image Retrieval (CBIR) system. However, there are a number of factors that are ignored when dealing with images which should be dealt with when using videos. These factors are primarily related to the temporal information available from a video document. Thus in this paper we begin with CBIR of Animal images and then extend it to the CBVR of Animal Videos.

2 CBIR of Animal Images

In this topic first we will discuss why traditional CBIR techniques cannot directly be used for CBIR of Animal images.

2.1 Problems of Using Color Histogram for CBIR of Animal Images

This technique cannot directly be used for Content based Animal image retrieval from Animal image Database. Two pictures shown below are of baby elephant and are almost similar.

Color histogram of these two images will be different. Thus if we compare these histogram we will get result that these two images are not similar, which is not true. That's why we cannot directly apply this method on the Animal image database.



Fig. 1.

2.2 Problems of Using Shape Descriptor for CBIR of Animal Images

This technique cannot directly be used for Content based Animal image retrieval from Animal image Database. Observe these two images as



Fig. 2.

Both the images are of same animal (i.e. tiger) but the shape descriptor for these two images will be different. Thus if one image is given as an input to the retrieval system, second image will not come in output. That's why we cannot use this method directly.

In above two approaches for CBIR the Animal (i.e. object) as well as its background was also taken into consideration while calculating the histogram (i.e. feature vector), which is different in every image. Thus these techniques are not performing well for Animals CBIR. Taking this into consideration paper [2] presents an efficient approach to CBIR of birds' images which can also be used for CBIR of Animal Images. Paper [2] presents a technique to extract an object (Animal) from the image by subtracting background of object (i.e. back ground of Animal) from image. After this it calculates color histogram of Animal object and uses it as a feature vector for CBIR.

Paper [1] presents a technique to calculate two key parameters WP and WA using color histogram of image. If these are used as feature vector to compare two images then it will speed up the CBIR process.

3 Antipole Tree

Antipole tree [3], [4] combine and extend ideas from the M-Tree, the Multi-vantage Point structure and the FQ-Tree to create a new structure in the “bisector tree” class, called the Antipole Tree. Bisection is based on the proximity to an “Antipole” pair of elements generated by a suitable linear randomized tournament. The final winners a, b of such tournaments are far enough apart to approximate the diameter of the splitting set. If $\text{dist}(a, b)$ is larger than the chosen cluster diameter threshold, then the cluster is split. This data structure is an indexing scheme suitable for (exact and approximate) best match searching on generic metric spaces. The Antipole Tree outperforms by a factor of approximately two over the existing structures such as List of Clusters, M-Trees, and others and, in many cases, it achieves better clustering properties.

4 CBVR for Animal Videos

In this topic first we will discuss the why traditional CBVR techniques are not suitable for CBVR of Animal videos.

4.1 Similarity of Videos

If a user inputs a query video Q and video V is very similar to query Q , meaning that video V contains most of shots similar to Q 's shots [5]. These shots will appear in the same order as they do in Q . That is, if V 's shots are not similar to Q 's or matching shots aren't in the same order, then the similarity between Q and V is low.

But this approach is not suitable for CBVR of Wild life animals, as two videos of the same animal may not have same shot sequence and probability of this is very high. Thus this conventional approach of CBVR is not suitable for CBVR of Wild life.

4.2 Proposed Approach for CBVR of Wild Life

As discussed above conventional Video segmentation processes is not suitable for CBVR of wild life. Thus we propose a new CBVR system for wild life.

Segmentation of Video

We have used the method described in Content Based Image Retrieval to compute the feature vector of frame in the video [1] [2]. Let video stream for segmentation is $S = \{F_1, F_2, F_3, F_4, \dots, F_n\}$ where, F_i is the frame in the video. Set F_1 as F_s .

Algorithm: SEGMENT_VIDEO (Video, Threshold)

Input: Video, threshold for selecting key frame

Output: Set of Key frame of videos

Steps

1. For each frame in video
 - a. Calculate feature vector (WP, WA) of the current frame
 - b. For each key frame identified till now in video
 - i. Calculate distance between key frame and current frame as

$$distance = \sqrt{((CF.WP - KF.WP) + (CF.WA - KF.WA))^2}$$

Where, CF = Current Frame and KF = Key Frame
 - ii. If (distance < threshold)
 1. Set Reject flag
 - c. If (! Reject Flag)
 - i. Add current frame to the Set of key frames
2. Add Feature vectors of Key frames to the video database
3. Stop

4.3 Searching in Video Database

To retrieve similar video(s) from the database, a query can be formed by a video or a sequence of images combined to form the video. Following Algorithm describes the searching process for video.

Algorithm: SEARCH_VIDEOS (Query Video, threshold for searching)

Input: Query Video, say q, threshold for searching, say threshold

Output: Similar video(s)

Steps

1. Create Antipole tree of feature vectors of key frame present in video database

2. Segment query video to identify key frame of the query video

Let $Q = \{F_{Q1}, F_{Q2}, \dots, F_{Qm}\}$ be the feature vectors of key frames of the query clip

Where, m is number of key frames of query video

3. For $I = 1$ to m ; every time increment I by 1
 - a. Search similar key frame Feature vectors for Feature vector F_{QI} in the Antipole Tree
 Let similar Key Frame feature vectors are $(SKF_1, SKF_2, \dots, SKF_p)$
 - b. Retrieve videos corresponding to those feature vectors.
 Let set of those videos be SV_I
4. Take the Union of sets
5. Stop

5 Experimental Results

All the experiments are performed on a Intel® CORE (TM) 2 Duo CPU with 1GB main memory, running Windows XP. The program is written in java and compiled with Net beans 6.0. Experiment database, which consists of 30 videos of different animals. Videos are downloaded from different websites.

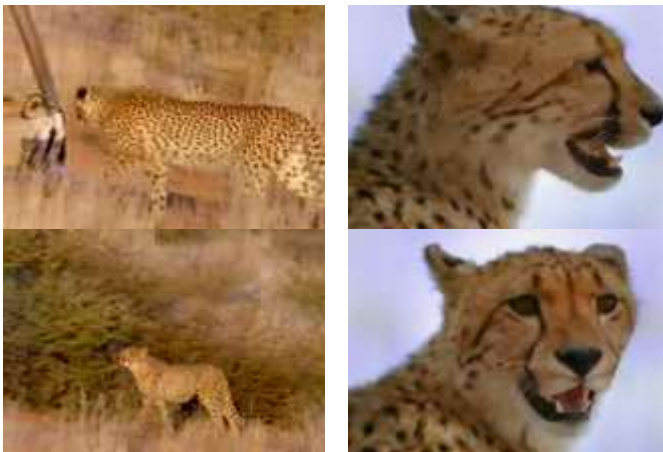


Fig. 3. Key Frames of Query Video

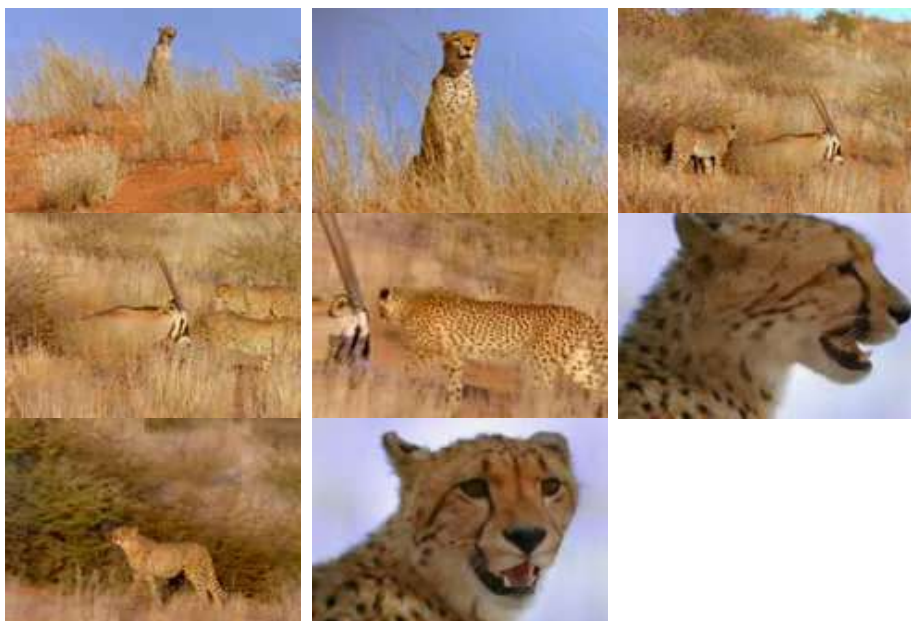


Fig. 4. Key Frames of Related Video 1

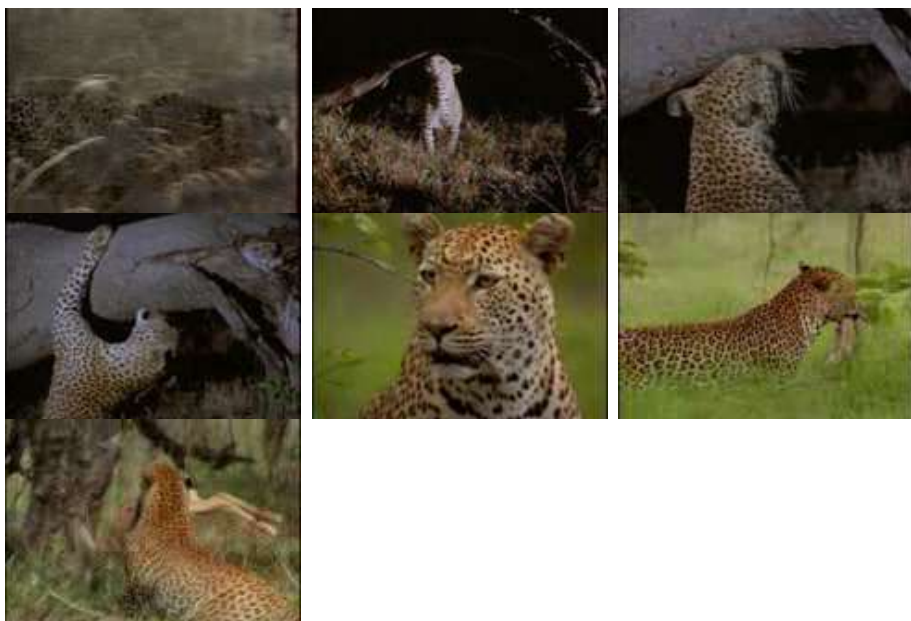


Fig. 5. Key Frames of Related Video 2

Video 1 and Video 2 are of same animal but are captured at different locations and at different time. Query video is sub part of Video 1. When we gave query video as input to the program it gave Video 1 and Video 2 as output.

6 Conclusion

In this paper, we have proposed an approach to content-based video retrieval of animal videos. Our proposed approach can deal with frame ordering and provides similarity retrieval. This approach can be very useful for Animal researches, they can search all most all the videos related to particular animal from Animal Video database.

References

1. Yihong, G., Hongjiang, Z., Chuan, C.H.: An Image Database System with Fast Image Indexing Capability Based on Color Histograms. In: IEEE Region 10's Ninth Annual International Conference TENCON 1994 (1994) ISBN: 0-7803-1862-5
2. Das, M., Manmatha, R.: Automatic Segmentation and Indexing in a Database of Bird Images. In: Proceedings of Eighth IEEE International Conference on Computer Vision, ICCV 2001 (2001) ISBN: 0-7695-1143-0
3. Cantone, D., Ferro, A., Pulvirenti, A., Recupero, D.R., Shasha, D.: Antipole Tree Indexing to Support Range Search and K-Nearest Neighbor Search in Metric Spaces. IEEE Transaction on Knowledge and Data Engineering 17(4) (April 2005) ISSN : 1041-4347
4. Battiato, S., Di Blasi, G., Reforgiato, D.: Advanced indexing schema for imaging applications: three case studies. IET Image Process. 1(3) (September 2007) ISSN : 1751-9659
5. Lee, A.J.T., Hong, R.-W., Chang, M.-F.: An Approach to Content-based Video Retrieval. In: 2004 IEEE International Conference on Multimedia and Expo., ICME (2004) ISBN: 0-7803-8603-5

Data Randomization for Synchronization in OFDM System

Rakhi Thakur¹ and Kavita Khare²

¹ Research Scholar

Department of EC, MANIT, Bhopal, India

² Associate Professor

Department of EC, MANIT, Bhopal, India

rakhi082003@yahoo.co.in,

kavita_khare1@yahoo.co.in

Abstract. Orthogonal Frequency Division Modulation (OFDM) is an efficient multi-carrier modulation scheme. This results in the optimal usage of bandwidth. It is used in many wireless communication systems due to its robustness towards fading channel behavior and a relative ease of implementation coming from computationally efficient Inverse Fast Fourier Transforms (IFFT). OFDM techniques form the basis of the physical layer of many broadband high data rate technologies including Digital Subscriber Lines (xDSL), Wi-Fi (IEEE802.11a/g/n) etc. One major problem with OFDM is synchronization between transmitter and receiver. This paper has been discussing the synchronization issues through data randomization.

Keywords: OFDM, LFSR, FEC, FFT, IFFT etc.

1 Introduction

Each sub-carrier in an OFDM system is modulated in amplitude and phase by the data bits. Depending on the kind of modulation technique that is being used, one or more bits are used to modulate each sub-carrier. The process of combining different sub-carriers to form a composite time-domain signal is achieved using Fast Fourier transform. Different coding schemes like block coding, convolutional coding or both are used to achieve better performance in low SNR conditions. Interleaving is done which involves assigning adjacent data bits to non-adjacent bits to avoid burst errors under highly selective fading. Block diagram of an OFDM transceiver is shown in fig 1.

2 Transmitter Pipeline

Transmitter control receives information from the MAC and generates the control and data for all the subsequent blocks. *Scrambler* randomizes the data bit stream to remove repeated patterns, like long sequences of zeros and ones. This enables better results for Forward Error Correction (FEC). A scrambler is usually implemented with linear feedback shift registers (LFSR). An LFSR has two algorithmic settings: the size of the shift register and the linear function, e.g $G(x) = X^{16} + X^5 + X^4 + X^3 + 1$ for generating the feedback.

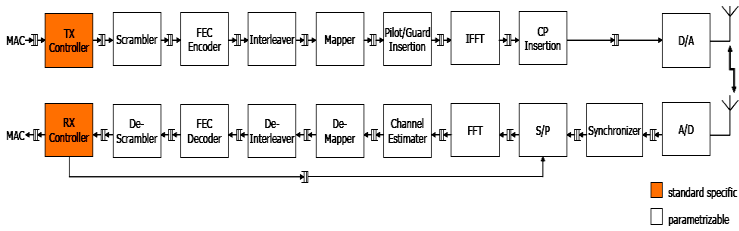


Fig. 1. Block diagram of the 802.11a OFDM transceiver

FEC Encoder encodes data and adds redundancy to the bit stream to enable the receiver to detect and correct errors. *Interleaver* rearranges blocks of data bits by mapping adjacent coded bits into non-adjacent sub carriers to protect against burst errors. *Mapper* passes interleaved data through a serial to parallel converter, mapping groups of bits to separate carriers, and encoding each bit group by frequency, amplitude, and phase. The output of the Mapper contains only the values of data subcarriers for an OFDM symbol. *Pilot/Guard Insertion* adds the values for pilot and guard subcarriers. *IFFT* converts symbols from the frequency domain to the time domain. *CP Insertion* copies some samples from the end of the symbol to the front to add some redundancy to the symbols. These duplicated samples are known as a cyclic prefix (CP). The purpose of the cyclic prefix is to avoid Inter-Symbol Interference (ISI) caused by multipath propagation. After CP insertion, the symbol are converted into analog signals by D/A converter and transmitted through the air[4].

3 Receiver Pipeline

The functionality of the blocks in the receiver is roughly the reverse of the functionality of their corresponding blocks in the transmitter. When the antenna detects the signal, it amplifies the signal and passes it to the A/D converter to generate baseband digital samples. *Synchronizer* detects the starting position of an incoming packet based on preambles. It is extremely important for the synchronizer to correctly estimate the OFDM symbol boundaries so that subsequent blocks process appropriate collection of samples together. *Serial to Parallel (S/P)* removes the cyclic prefix (CP) and then aggregates samples into symbols before passing them to the FFT. *FFT* converts OFDM symbols from the time domain back into the frequency domain. *Channel Estimator* estimates and corrects the errors caused by multipath interference. *Demapper* demodulates data and converts samples to encoded bits, which are used by the FEC decoder. *De-interleaver* reverses the interleaving performed by transmitter and restores the original arrangement of bits, *FEC Decoder* uses the redundant information that was introduced at the transmitter to detect and correct any errors that may have occurred during transmission. *Descrambler* reverses the scrambling performed by the transmitter. *RX Controller* based on the decoded data received from Descrambler, the RX Controller generates the control feedback to S/P block[4].

4 Functional Description of Scrambler

The addition of components to the original signal or the changing of some important component of the original signal in order to make extraction of the original signal difficult accomplishes scrambling. In telecommunications and recording, a scrambler (also referred to as a randomizer) is a device that manipulates a data stream before transmitting. A descrambler at the receiving side reverses the manipulations. A scrambler can be placed just before a FEC coder, or it can be placed after the FEC, just before the modulation or line code. A scrambler replaces sequences into other sequences without removing undesirable sequences, and as a result it changes the probability of occurrence of vexatious sequences. There are two main reasons why scrambling is used[1]:

- A. It facilitates the work of a timing recovery circuit (see also Clock recovery), an automatic gain control and other adaptive circuits of the receiver (eliminating long sequences consisting of '0' or '1' only).
- B. It eliminates the dependence of a signal's power spectrum upon the actual transmitted data, making it more dispersed to meet maximum power spectral density requirements (because if the power is concentrated in a narrow frequency band, it can interfere with adjacent channels).

4.1 Types of Scramblers

1. Additive (synchronous) scramblers
2. Multiplicative(self-synchronizing) scramblers

Multiplicative scramblers (also known as feed-through) are called so because they perform a multiplication of the input signal by the scrambler's transfer function in Z-space. They are discrete linear time-invariant systems. A multiplicative scrambler is recursive and a multiplicative descrambler is non-recursive. Unlike additive scramblers, multiplicative scramblers do not need the frame synchronization, which is why they are also called self-synchronizing. The additive descrambler is just the same device as the additive scrambler. Additive scrambler/descrambler is defined by the polynomial of its LFSR (for the scrambler on the picture below, it is $(1 + x^{-3} + x^{-4} + x^{-5} + x^{-16})$ and its initial state[3].

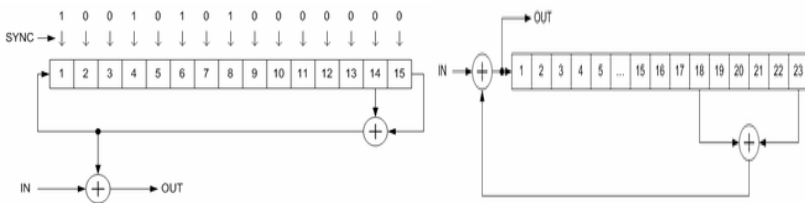


Fig. 2. A)An Additive Scrambler B) Multiplicative (self-synchronizing) scramblers

5 Fpga Implementation

The LFSR implements the following polynomial: $G(x) = X^{16} + X^5 + X^4 + X^3 + 1$ ----(1)

Traditionally the LFSR is clocked at the bit transfer rate. The output of the LFSR is XORed with the data to form the scrambled data. With a data bit rate of 2.0Gbs, serially scrambling the data is not very practical. A method for implementing the LFSR in parallel needs to be developed to make any design practical. The most intuitive method would be to scramble the data 8-bits at a time. This effectively allows the designer to work with a byte rate of 250 MHz rather than the bit rate of 2Ghz.

Scrambling 8-bits at a time in parallel is straight forward, and easy to implement, but the operating frequency of 250 MHz is still rather high and might be hard to work with if designers are using FPGA's. Implementing the PCI Express scrambler/de-scrambler in a 16-bit parallel fashion would essentially halve the frequency designers would be required to work with (125 MHz). This would greatly help designers to meet timing much easier when working with FPGA's[5]. The implementation of the PCI Express Polynomial in parallel 16-bits at a time is not as trivial as the 8-bit parallel implementation and thus causes it to have greater value.

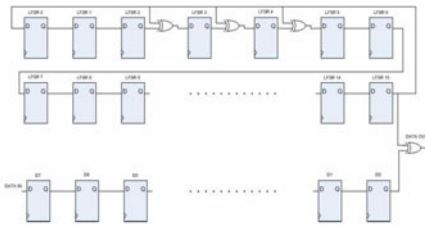


Fig. 3. Serial Scrambler

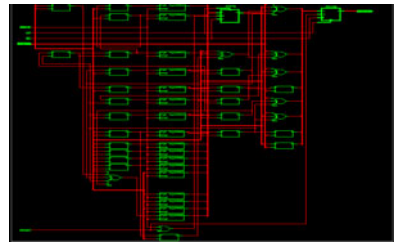


Fig. 4. RTL view of parallel scrambler

A separated synchronizing scrambler/ descrambler pair that removes the possibility of catastrophic error due to improper transmission of initial condition information without disrupting the OFDM modulation. A transmitting device within the pair includes a first and a second data scrambler wherein the first data scrambler couples to receive the incoming data stream and filters the incoming data stream to provide a first filtered signal using a key signal. The second data scrambler, having an initial condition, couples to receive the first filtered signal and converts it into a scrambled signal using a scrambling seed. The second data scrambler comprises a random series generator for generating the scrambling seed to convert the first filtered signal into a scrambled signal. The scrambled signal is transmitted to the receiving device. A receiving device within the pair includes a first and a second data descrambler coupled together. The first data descrambler, having an initial condition equivalent to that of the second data scrambler, couples to receive the transmitted scrambled input data stream to convert it into a descrambled signal using a descrambling seed equivalent to the scrambling seed. The first data descrambler includes a random series generator for generating the descrambling seed. The second data descrambler couples

to receive the descrambled signal and filters it to provide a filtered descrambled signal using a feed forward filter and a key signal. Scramblers have certain drawbacks[2]:

- 1) Both types may fail to generate random sequences under worst-case input conditions.
- 2) Multiplicative scramblers lead to error multiplication during descrambling (i.e. a single bit error at the descrambler's input will result into w errors at its output, where w equals the number of the scrambler's feedback taps).
- 3) Additive scramblers must be reset by the frame sync; if this fails massive error propagation will result as a complete frame cannot be descrambled.

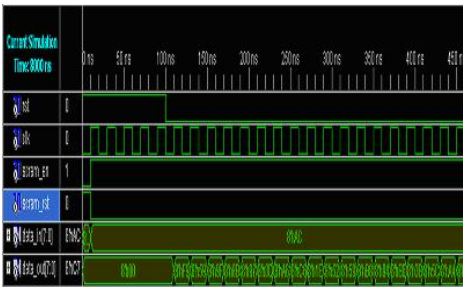


Fig. 5. Simulation of parallel scrambler

Device Utilization Summary			
Logic Utilization	Used	Available	Utilization
Number of Slice Flip Flops	25	9,312	1%
Number of 4 input LUTs	42	9,312	1%
Logic Distribution			
Number of occupied Slices	22	4,656	1%
Number of Slices containing only related logic	22	22	100%
Number of Slices containing unrelated logic	0	22	0%
Total Number of 4 input LUTs	42	9,312	1%
Number of bonded I/Os	20	92	21%
Number of BUFGMUXs	1	24	4%

Fig. 6. Design Summary

6 Conclusion

The OFDM implementation provides the basic properties required to both transmit and receive data through stationary channels. The system corrects for multipath, frequency and timing offsets, and can transfer data with a variable number of subcarriers, cyclic prefix length, and subcarrier modulation. There are more properties that we are currently making progress on implementing. Scrambler provides synchronization between transmitter and receiver. From Design summary it is to be noted that it uses small area and operating frequency as well as speed is higher.

References

1. Carlos Colderon, J., Tapia, J.M.: VHDL modeling and synthesis of scrambler and CRC processor for B-ISDN, barcelona Spain
2. Ng, M.C., Vijayaraghavan, M., Dave, N., Arvind, Raghavan, G., Hicks, J.: Techniques for High-Level IP Reuse across Different OFDM Protocols. Nokia Research Center Cambridge Nokia Corporation
3. Rondeau, T.W.: CTVR, Trinity College Dublin, Dublin, Ireland, Proceedings of the SDR 2008 Technical Conference and product Exposition, Copyright © 2008 SDR Forum (2008)
4. Pandey, A., Agrawalla, S.R., Manivannan, S.: VLSI implementation of OFDM Modem. Wipro Technology
5. European Journal of Scientific Research 31(3), 319–328 (2009) ISSN 1450-216X
6. Xilinx Inc., Spartan3E Starter Kit Board User Guide

Split Step Method in the Analysis and Modeling of Optical Fiber Communication System

Saroja V. Siddamal¹, R.M. Banakar¹, and B.C. Jinaga²

¹ Dept. of EC, BVBCET, Hubli, India
{sarojavs, banakar}@bvb.edu

² Dept. of EC, JNTU, Hyderabad, India
jinaga@yahoo.co.in

Abstract. The increasing complexity of optical processing algorithm has led to the need of developing the algorithms specifications using software implementation that became in practice the reference implementation. Adapting the algorithm specified by such software models into architectures becomes a very resource consuming and memory intensive task. The key objective in this paper is to develop analytical models to analyze the effects of various parameters such as propagation distance, chirping factor on the received power at the receiver end. SSF algorithm is simulated using MATLABTM on Intel[R]TM Core[TM] 2 Duo CPU T5470 @ 1.60GHz, 0.99GB of RAM. Analysis show the received output power reduces by 61.1% with 5 time increase in chirping factor. Increase in chirp decreases the received power and broadens the propagation pulse. A reduction of 82% of received power is observed as the propagation distance increases by 50Kms. Mapping of SSFM algorithm to architecture is done with this analysis. The algorithm is synthesized using Xilinx design Manager. The hardware is implemented using virtex XC5VLX30TFF655 FPGA device family with speed grade -6. The complete hardware operates with maximum frequency 20.982Mhz which uses total memory of 523564Kbytes.

Keywords: Nonlinear Schrodinger equation, Split Step Fourier Transform, dispersion, fiber nonlinearities, chirp.

1 Introduction

Rapid advances in optical communications technology has paved the pathway for high-speed high-capacity transmission of signal and data over long distances, enabling novel applications such as on-demand television. Optical communications systems uses a very cost effective technology called the Wavelength division multiplexing (WDM) which efficiently exploits the bandwidth of the optical fiber by simultaneously transmitting multiple channels in non-overlapping, closely spaced wavelengths. However, due the presence of many physical impairments such as chromatic dispersion, polarization mode dispersion, cross-phase modulation, self-phase modulation, four-wave mixing and amplified spontaneous emission due to loss compensating optical amplifiers, optical fibers cannot be just deployed for

building networks. Because almost all the data sent through fibers are digital and hence inevitably in pulsed form, the influence of the above impairments on transmitted pulses need to be fully understood to design networks. Researchers have found that Nonlinear Schrodinger Equation (NLSE) mapped to optical domain provides an adequate framework to describe optical signal propagation through nonlinear fiber for all the engineering applications found in practice. The most commonly used method in the fiber span simulation is split step method (SSM)[\[4\]](#), which is based on the partition of the fiber spans into several spatial steps. Over each step the linear (dispersion) and nonlinear operators (Kerr Effect) of the NLSE are considered separately therefore these phenomena are supposed to act independently to each other. The numerical solution of NLSE through SSM converge to the exact solution when the split step tends to zero[\[4\]](#). But since smaller is the step size higher is the computational time and a trade-off between these two opposite requirements must be found out. The rest of the paper is organized as follows, section 2 reviews wave propagation in optical fiber. Analysis of fiber nonlinearities is done using NLSE in section 3. Section 4 pave the way to SSFM architecture. Section 5 discusses simulation result followed by RTL validation and conclusion.

2 Wave Propagation in Optical Fiber

There are two primary system parameters which determine the characteristics of optical communication systems. Specifically, data is transmitted by a sequence of pulses, and the system must ensure these pulses are received with a sufficiently low probability of error, also called the bit-error rate (BER). Given a particular receiver, achieving a specified BER requires a minimum received power and a maximum data rate or signal bandwidth. An optical fiber introduces attenuation and dispersion in the system. Whereas attenuation tends to increase the power requirements of the transmitter needed to meet the power requirements at the receiver, dispersion limits the bandwidth of the data which may be transmitted over the fiber. Dispersion refers to the distortion of a propagating wave. In this paper we have analyzed the effects of chirp and propagation distance on the received optical power. These results are used to develop the hardware to analysis the effects of dispersion and nonlinearity on optical transmission.

2.1 Effect of Chirp

Chirp parameter estimation is a well-known problem in signal processing community. Chirp signals occur in many applications, e.g., radar, sonar, bioengineering, gravity waves and seismograph. A pulse can acquire a chirp e.g. during propagation in a transparent medium due to the effects of chromatic dispersion and nonlinearities (e.g. self-phase modulation arising from the Kerr effect). In semiconductor lasers or amplifiers, chirps can also result from refractive index changes associated with changes in the carrier density. The chirp of a pulse can be removed or reversed by propagating it through optical components with suitable chromatic dispersion.

2.2 Nonlinear Effect

The nonlinear effect in optical fiber occur either due to intensity dependence of refractive index of the medium or due to inelastic scattering phenomenon. The power dependence of the refractive index is responsible for Kerr-effect. Depending upon the type of input signal, the kerr-nonlinearity manifests itself in three different effects such as self-phase modulation, cross-phase modulation and four waves mixing. At high power level, the inelastic scattering phenomenon can induce stimulated effects such as stimulated Brillouin Scattering and stimulated raman scattering. The intensity of scattered light grows exponentially if the incident power exceeds a certain threshold value. The nonlinearity effect depends on transmission length longer the fiber link length, the more the light interaction and the nonlinear effects. To fully utilize the fiber bandwidth, numerous channels a different wavelength can be multiplied on a single optical fiber known as wavelength division multiplexing. Problems appear in wavelength division multiplexing system as lengthy fiber accumulate high dispersion and multiple channels introduce a large amount of optical power, which ultimately enhance the fiber nonlinearity.

3 Analysis of Fiber Nonlinearities by Nonlinear Schrodinger Equation

3.1 Nonlinear Schrodinger Equation (NLSE)

Within the slowly varying amplitude approximation of Maxwell's equations, signals propagating through optical fiber is described using the equation 1

$$\frac{\partial A(z, \tau)}{\partial z} = -\frac{j}{2}\beta_2 \frac{\partial^2 A(z, \tau)}{\partial \tau^2} + \frac{\beta_3}{6} \frac{\partial^3 A(z, \tau)}{\partial \tau^3} + j\gamma |A(z, \tau)|^2 A(z, \tau) - \frac{\alpha}{2} A(z, \tau) \quad (1)$$

where $j = \sqrt{-1}$, z is the distance measured along the fiber, τ is the time in a inertial frame with velocity equal to the phase velocity of the fiber medium, $A(z, \tau)$ is the slow varying amplitude of the optical field, β_2 is second order dispersion coefficient, β_3 is the third order dispersion coefficient, α is the attenuation coefficient and γ is the nonlinear coefficient of fiber. The dispersion parameters lead to broadening of pulses due to having different group velocities of pulse spectral components at different frequencies. The nonlinearity leads to the distortion of pulse depending on its power content. The above NLSE equation can be written using the following operator notation

$$\frac{\partial A(z, \tau)}{\partial z} = (\mathcal{L} + \mathcal{N}) A(z, \tau) \quad (2)$$

where \mathcal{L} and \mathcal{N} are linear and nonlinear operators with following definitions:

$$\mathcal{L} \triangleq -\frac{j}{2}\beta_2 \frac{\partial^2}{\partial \tau^2} + \frac{\beta_3}{6} \frac{\partial^3}{\partial \tau^3} - \frac{\alpha}{2}$$

$$\mathcal{N} \triangleq j\gamma |A(z, \tau)|^2$$

The main advantage of this splitting is that both the linear and the nonlinear part have analytical solutions valid for the entire parameter range of interest even though such general analytical solutions can be found for the original NLSE. In SSFM, these analytical solutions valid for linear and nonlinear parts used to construct the solution of the original NLSE.

3.2 Split Step Fourier Method (SSFM)

SSFM splits the nonlinear Schrodinger equation into linear and non-linear parts. It accurately models the fiber nonlinearities within the system. The nonlinear effects and dispersion work in conjunction within transmission fiber. As SSFM is only approximation method, these two factors can be broken up and solved individually. The simulation step length, the fiber nonlinear coefficient γ , dispersion parameter β , system parameters and the input test vector (Gaussian pulse) are the inputs to the SSFM software model. The numerical SSFM System parameters used during functionality testing are listed in Table I. Split step Fourier method is a very effective and efficient numerical method to solve NLSE under general conditions. In SSFM the fiber is divided into relatively small steps of equal length h such that the operators \mathcal{L} and \mathcal{N} commute with each other approximately within such a segment. Under such conditions, the solution to the NLSE at step $z = kh$, where k is an integer is given by

$$A(kh, \tau) \approx \exp(\mathcal{L}h + \mathcal{N}h)A((k - 1)h, \tau) = \exp(\mathcal{L}\frac{h}{2})\exp(\mathcal{N}h)\exp(\mathcal{L}\frac{h}{2})A((k - 1)h, \tau) \quad (3)$$

where $(A(k - 1)h, \tau)$ is the field solution for the previous step. The corresponding numerical implementation of equation 3 can be written as:

Table 1. SSFM System Parameters

Initial peak power P_o	21 mw
Chirping parameter C_o	1
Gaussian pulse order M	1
Initial pulse width t_o	5×10^{-12}
Nonlinear coefficient N	$0.0024 \text{ km mw}^{-1}$
Speed of light	$3 \times 10^{-5} \text{ km s}^{-1}$
Propagation distance Z_{final}	100 km
Simulation step length	4 km
Partition of sections M	125

$$A(kh, \tau) = \mathbf{IFFT}(\mathcal{F}(\exp(\mathcal{L}\frac{h}{2})))\mathbf{FFT}(\exp(\mathcal{N}h))\mathbf{IFFT}(\mathcal{F}(\exp(\mathcal{L}\frac{h}{2})))\mathbf{FFT}(A((k - 1)h, \tau)) \quad (4)$$

where \mathcal{F} is the Fourier transform operator and **FFT** and **IFFT** are fast Fourier transform and its inverse respectively. A schematic diagram illustrating this one

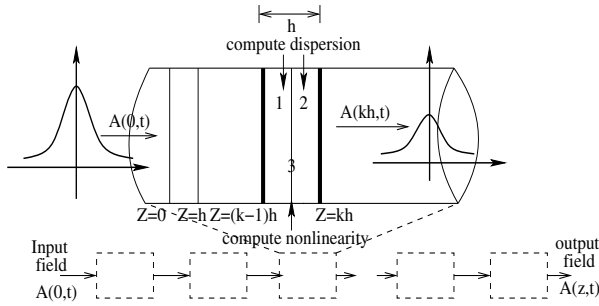


Fig. 1. Schematic for one step of SSFM

simulation step length implementation of SSFM is illustrated in Fig. 1. It clearly shows the three steps in the implementation of SSFM. In the first and the second part i.e point 1 and 2 dispersion is computed. In the centre of step at point 3 nonlinearity is computed. The above steps are repeated for the entire length of the fiber to obtain the final field solution $A(L, t)$ where L is the length of the fiber. In the SSFM model used maximum allowable tolerance, two iterations are utilized to increase the accuracy in nonlinearity and dispersion computations.

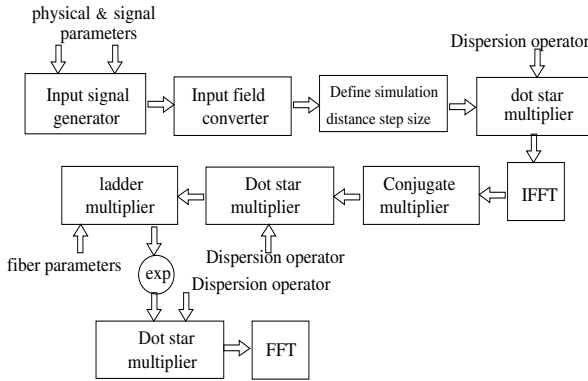


Fig. 2. Datapath of SSFM architecture for dispersion computation for first half propagation distance

4 SSFM Architecture

The data-flow/signal flow corresponding to the SSF implementation is as shown in fig 2. The MATLABTM [12] program has been utilized for functionality testing, parameter selection and input-output test data repository. The algorithm is divided into three sub-computational stages.

1. Dispersion Computation Stage 1 (DCS1): This stage utilizes the dispersion parameter specifications from the fiber parameters set values. DCS1 is responsible to compute the dispersion value set using a multiplication step and then to time domain set values using IFFT.
2. Nonlinearity Computation Stage (NCS): This stage uses the nonlinearity parameters specifications from the fiber parameter set. NCS assesses the nonlinearity in the SSFM model according to the input parameters and then the data is converted to frequency domain for another DCS.
3. Dispersion Computation Stage 2 (DCS2): As depicted in Fig. 1, SSFM dispersion is again calculated for the next half propagation distance.

Fig 2 shows the datapath for dispersion estimation during first half propagation distance.

5 Simulation and Results

The SSFM algorithm is simulated using MATLABTMCore[TM] 2 Duo CPU T5470 @ 1.60GHz, 0.99GB of RAM. A single Gaussian pulse is assumed as the input data with suitable initial peak power of signal source, chirping parameter and initial pulse width to the matlab code. The analysis for various parameters like effect of chirping factor, propagation distance on the received power is performed.

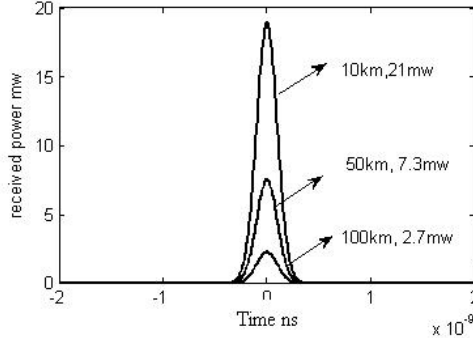


Fig. 3. Received power for various propagation distances

5.1 Effect of Propagation Distance on Received Power

Fig. 3 shows the variation of received power due to change in propagation distance. For a propagation distance of 5Km the received the power is 23mw considering the chirp $C=0$. With the increase in the propagation distance to 50Km the received power reduces to 7.3mw. A reduction of 81% of received power is observed as the propagation distance increases by 50kms. Hence the propagation distance becomes the key factor to estimate the fiber dispersion and nonlinearity.

5.2 Effect of Chirping Factor on Received Power

Fig. 4 shows the variation of received power with respect to time for various the chirping factor. For the chirping factor $C=0$ the received power is 21mw and the width of the pulse is 1.5×10^{-9} ns. As the chirping factor increases to $C=5$, the power received is 8mw and the width of the received pulse is 3×10^{-9} ns. The experimental results show that as the chirping factor increases the pulse broadens. This boarding of pulse results in the reduction of received power at the output.

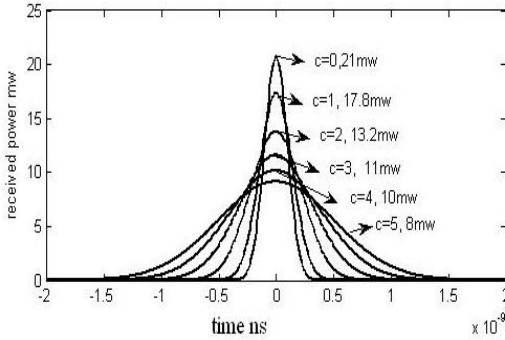


Fig. 4. Received power for various chirping factors

5.3 RTL Validation

A mapping of SSFM from algorithm to architecture is done considering the above analysis. SSFM algorithm for an input of 1024 is synthesized using Xilinx design Manager. The algorithm is implemented using virtex XC5VLX30TFF655 FPGA device family with speed grade -6. The hardware model is analysed for the recived power for the propogation of $5Km$ and 0 chirping factor. The maximum frequency of operation is $20.982Mhz$. The total memory usage is $523564Kbytes$.

6 Conclusion

Fiber nonlinearities have become one of most significant limiting factors of system performance since the advent of erbium-doped fiber amplifiers (EDFAs) because input power is increasing and the effects of fiber nonlinearities are accumulating with the use of EDFAs. In wavelength-division-multiplexing (WDM) systems, inter- channel Interference due to fiber nonlinearities may limit the system performance significantly. Therefore, understanding of fiber nonlinearities is crucial to optimize system performance of optical fiber transmission. With the understanding of the simulation results the SSFM algorithm is mapped to architecture. SSFM for Gaussian input is synthesized using Xilinx design manager. The algorithm is implemented using virtex XC5VLX30TFF655 FPGA device family with speed grade -6.

References

1. Agrawal, G.P.: *Fiber-Optic Communication Systems*, 3rd edn. Wiley Interscience, Hoboken (2002)
2. Mukherjee, B.: *Optical WDM Networks*. Springer, Heidelberg (2006)
3. Agrawal, G.P.: *Nonlinear Fiber Optics*, 4th edn. Academic Press, London (2006)
4. Zoldi, S., Ruban, V., Zenchuk, A., Burtsev, S.: Parallel Implementation of the Split-step Fourier Method For Solving Nonlinear Schrödinger. From *SIAM News* 32(1)
5. Liu, X., Lee, B.: A Fast Methods for Nonlinear Schrödinger Equation. *IEEE Photonics Technology Letters* 15(11) (November 2003)
6. Weideman, J.A.C., Herbst, B.M.: Split-step methods for the solution of the nonlinear Schrödinger equation. *SIAM J. Numer. Anal.* 23, 485–507 (1986)
7. Harboel, P.B., Souza, J.R.: Assessment of higher-order exponential operators for the simulation of high capacity optical communication systems by the split-step fourier method. *Journal of Microwaves and Optoelectronics* 3(2) (August 2003)
8. Jiang, Y., Zhou, T., Tang, Y., Wang, Y.: Twiddle-Factor-Based FFT Algorithm with Reduced Memory Access. In: *Proceedings of the International Parallel and Distributed Processing Symposium, IPDPS 2002* (2002)
9. Zoldi, S.M., Ruban, V., Zenchuk, A., Burtsev, S.: Parallel Implementations of the Split-Step Fourier Method for Solving Nonlinear Schrödinger Systems. Los Alamos National Laboratory (May 29, 2006)
10. Sinkin, O.V., Holzlohner, R., Zweck, J., Menyuk, C.R.: Optimization of the Split-Step Fourier Method in Modeling Optical-Fiber Communications Systems. *Journal of Light Wave Technology* 21(1) (January 2003)
11. Menyuk, C.R., Carter, G.M., Kath, W.L., Mu, R.-M.: Dispersion managed solitons and chirped return to zero: What is the difference? In: Kaminow, I.P., Li, T. (eds.) *Optical Fiber Telecommunications*, ch. 7, pp. 305–328. Academic, San Diego (2002)
12. Jong-Hyung: Analysis and Characterization of fiber Nonlinearities for deterministic and stochastic signal sources
13. Siddamal, S.V., Banakar, R.M., Jiniga, B.C.: Optimization of split step Fourier method using reduced factor in modeling fiber Optics Communication. In: *International Conference 3CI Bangalore* (2007)

Performance Analysis of MIMO- Space-Time Trellis Code System over Fading Channels

Sandeep Bhad¹ and A.S. Hiwale²

¹ Disha Institute of Management and Technology, Raipur (CG), India

² Genba Sopanrao Moze College of Engineering, Balewadi, Pune (MS), India
sandeepbhad@gmail.com, ashiwale@gmail.com

Abstract. This paper discusses the error performance of Space- Time Trellis Code system over fading channels. Multiple-input multiple-output (MIMO) technology constitutes a breakthrough in the design of wireless communications systems, and is already at the core of several wireless standards. It offers high data rate and excellent throughput of wireless communication system. Trellis-code modulation (TCM) is one of the bandwidth efficient coding modulation techniques used in digital communications system. Here we discuss the design criteria for fading channels, employs that transmit and receive antenna diversity. It is observed from the simulation results that the code performance is improved by increasing the number of states. Also it provided the coding advantage as the number of states and receives antennas increases. The simulation have been conducted for 4 Phase Shift Keying (PSK) with data rate 2 b/Hz/s for two transmit and two receive antennas ($N_T=N_R=2$) system, provided better performance using 64 states STTC encoder.

Keywords: MIMO, TCM, Space-Time Trellis Code (STTC), Fading Channels, Code Design Criteria.

1 Introduction

The Multi antenna and Space time coding system involved into a most vibrant research area in wireless communication. MIMO techniques deliver significant performance enhancements in terms of high data transmission rate and interference reduction. Space - Time coding is a powerful technique that improves the error performance of multi-antenna wireless communications systems by introducing spatial-temporal correlation between transmitted code words. As compare to other ordinary channel codes, space time code promises protection against channel fading, noise, and interference. A space time trellis code allow serial transmission of symbols with multi antenna system.[10] In the first performance investigation of the space-time trellis codes (STTCs) [1], analytical bounds and design criteria were proposed for fading channels. It was summarized that in fading channels, the critical parameters of rank, determinant & trace criteria. Based on these criteria, new 4-phase shift keying (PSK) STTCs have been reported in [7] slow fading channels, and in [8] for fast fading channels.

The theme of this paper is the MIMO- space time trellis code (STTC) performance analysis on the basis of different code design criteria over fading channels. We are essentially interested in quantifying the impact of multi antenna system with increasing the number of states (4, 8, 16, 32 and 64 states) on diversity order and coding gain provided by space time code.

2 MIMO System Model

Fig. 1 illustrates a generic block diagram of MIMO system with N number of transmit antennas and M number of receive antennas. The communication channel is assumed to be frequency flat Rayleigh fading channel. Over T symbol duration, a codeword C of size $N_T \times T$ is transmitted through N_T transmit antennas. The channel is described by an $N_R \times N_T$ complex matrix, denoted by H . The ij^{th} component of the matrix H , denoted by h_{ij} , represents the channel fading coefficient from the j^{th} transmit to the i^{th} receive antenna for normalization purposes we assume that the received power for each of M receives branches is equal to the total transmitted power. Physically, it means that we ignore signal attenuations and amplifications in the propagation process, including shadowing, antenna gains etc. Thus we obtain the normalization constraint for the elements of H , on a channel with fixed coefficients, as

$$\sum_{j=1}^{N_T} |h_{i,j}|^2 = N_T, i=1,2,\dots,N_R \tag{1}$$

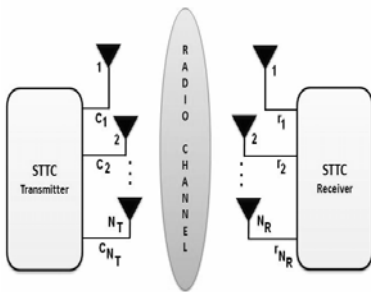


Fig. 1. MIMO System Model

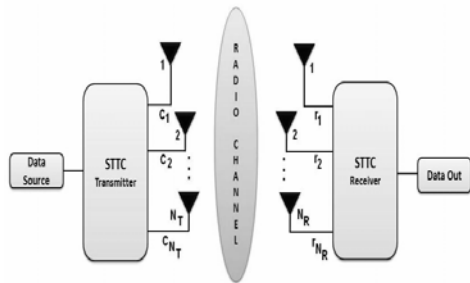


Fig. 2. MIMO STTC System Model

3 Error Performance of STTC MIMO

Space Time Trellis Coding (STTC) is a coding technique for MIMO that provides full diversity as well as coding gain [2]. STTC combines coding, modulation, transmit diversity and optional receive diversity. The coding gain is obtained at the cost of increased decoding complexity. Fig. 2 shows a block diagram of a MIMO-STTC, with N_T transmit antennas and N_R receive antennas. First, the channel code encodes

the source data. The space-time trellis encoder maps one symbol at a time, to a $N_T \times 1$ vector output. The channel code creates correlation between codewords across time (between successive symbols) and space (between different transmit antennas). At the receiver, the implemented STTC decoder is the Viterbi decoder. As the number of states increase, the FER performance improves at the cost of additional decoding complexity.

Early works by Tarokh et al. [03] provided theoretical background on space-time codes. Tarokh *et al.* developed the rank and determinant criteria for the quasi-static flat fading channel. The rank criterion may lead to full diversity, and makes Tarokh's criteria extremely useful for MIMO systems. Rank and determinant design criteria have been widely viewed as the best way to code design. The weakness of Tarokh's criteria is that the code design criteria are based on minimizing the worst upper bound of pairwise error probability (PEP). Essentially, these criteria are also based on the assumption that the error events with minimum determinant are dominant. Although this is true for AWGN channel, this is not always the case for the quasi-static flat fading channel where there are no dominant error events [4]. Thus, although the *rank criterion* is of great importance, the determinant criterion does not provide guidelines for optimal coding gain.

In this paper it is shown that the performance of the 4 states space-time trellis codes proposed is superior to that of some existing codes for two transmit antennas and two receive antenna.

Let us assume that each element of the signal constellation is contracted by a scale factor $\sqrt{E_s}$ chosen so that the average energy of the constellation elements is 1. Thus the design criteria given in section 4, is not constellation-dependent and applies equally well to 4-PSK, 8-PSK, and 16-QAM.

The probability that a maximum-likelihood receiver decides erroneously is given by

$$e = e_1^1 e_1^2 \dots e_1^{N_T} e_2^1 e_2^2 \dots e_2^{N_T} \dots e_l^1 e_l^2 \dots e_l^{N_T} \tag{2}$$

assuming that

$$c = c_1^1 c_1^2 \dots c_1^{N_T} c_2^1 c_2^2 \dots c_2^{N_T} \dots c_l^1 c_l^2 \dots c_l^{N_T} \tag{3}$$

was transmitted,

Assuming ideal channel state information (CSI) the probability of transmitting C and deciding in favor e at the decoder is approximated by [2]

$$P(c \rightarrow e | h_{i,j}, i=1,2,\dots,N_T, j=1,2,\dots,N_R) \leq \exp(-d^2(c, e) E_s / 4N_o) \tag{4}$$

where $N_o/2$ is the noise variance per dimension and

$$d^2(c, e) = \sum_{j=1}^{N_R} \sum_{t=1}^l \left| \sum_{i=1}^{N_T} h_{i,j} (c_t^i - e_t^i) \right|^2 \tag{5}$$

This is just a standard approximation to the Gaussian tail function [2]. In case of Rayleigh fading, an upper bound on the average probability of error can be computed by

$$P(c \rightarrow e) \leq \left(\frac{1}{\prod_{i=1}^{N_T} (1 + \lambda_i E_s / 4 N_o)} \right)^{N_R} \quad (6)$$

Let r denote the rank of matrix, then the kernel of matrix has dimension $N_T - r$ and exactly $N_T - r$ eigen values of matrix are zero. Let the non zero eigen values of matrix are $\lambda_1, \lambda_2, \dots, \lambda_r$, then it follows from inequality of eq. (6) that

$$P(c \rightarrow e) \leq \left(\prod_{i=1}^r \lambda_i \right)^{-N_R} (E_s / 4 N_o)^{-r N_R} \quad (7)$$

Also the probability error is given by [6]

$$P(c, e) \leq 1 / 4 \exp \left(-N_R \frac{E_s}{4 N_o} \sum_{i=1}^r \lambda_i \right) \quad (8)$$

Thus the diversity advantage of $N_R r$ and coding advantage of $(\lambda_1 \lambda_2 \dots \lambda_r)^{1/r}$ is achieved.

4 Performance of the Code Based on Different Design Criteria

This section discusses the criteria design for space time trellis code system. The code frame error rate (FER) performance is evaluated by simulation using different states. In general, the rank and determinant criteria are used for single receive antenna where trace criterion is use for two and more than two receive antennas. The performance of the code is improved by increasing the number of states. Maximum likelihood viterbi algorithm provides the lowest path computation.

4.1 Design Criteria for Fading Channels

The error performance upper bounds given by eq. (6) and eq.(7) indicate that the design criteria for slow Rayleigh fading channels will [4] depend on the value of $r N_R$. The maximum possible values of $r N_R$ is $N_T N_R$ when $N_T = N_R$. For small values of $N_T N_R$, corresponding to a small number of independent sub channels, the error probability is dominated by the minimum rank r of channel matrix over all possible codeword pairs. The product of the minimum rank and the number of receive antennas, $r N_R$ is called the minimum diversity. In addition, in order to minimize the error probability, the minimum product of the nonzero eigen values of matrix along

the pairs of codeword with the minimum rank should be maximized. Therefore, if the value of $N_T N_R$ is small, the STC design criteria for slow Rayleigh fading channels can be summarized as given below [5][12].

Design Criteria Set 1:

- 1) Maximize the minimum rank r of matrix over all pairs of distinct codewords.
- 2) Maximize the minimum Product $\prod_{i=1}^r \lambda_i$ of matrix along the pairs of distinct codewords with the minimum rank.

This set of design criteria is referred to as the rank and determinant criteria. For large values of $N_T N_R$ corresponding to a large number of independent sub channels, the PEP is upper bounded. In order to get an insight into the code design for systems of practical interest, we assume that the STC operates at a reasonably high SNR, which corresponds to

$$\frac{E_S}{4N_O} \geq \frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^r \lambda_i^2} \quad (9)$$

Design Criteria Set 2:

- 1) Make sure that the minimum rank of r matrix over all pairs of distinct codewords is large enough, such that $r N_R \geq 4$
- 2) Maximize the minimum trace $\sum_{i=1}^r \lambda_i$ of matrix among all pairs of distinct codewords. It is important to note that this design is consistent with the trellis-code design in fading channels with a large number of diversity branches which introduced reduces the effect of fading, the channel approaches an AWGN model [12]. Therefore, the trellis-code design criteria for AWGN channels apply to fading channels with a large number of diversity. In a similar way, in STC design, when the number of independent sub channels $r N_R$ is large, the channel converges to an AWGN channel. Thus, the code design is the same as that for AWGN channels.

In general, for random variables with smooth pdfs, the central limit theorem can be applied if the number of random variables in the sum is larger than four [9]. In the application of the central limit theorem, the choice of four as the boundary has been further justified by the code design and performance simulation, as it was found that as long as $r N_R \geq 4$, the best codes based on the trace criterion outperform the best codes based on the rank and determinant criteria [3][5].

5 Simulation Results

On the basis of above mention methodology code performance of the MIMO-STTC system over fading channels are presented. The simulated communication system had two transmit and two receive antennas (i.e $N_T=2$, $N_R=2$). The source symbols were transmitted with 130 frames, 2 b/s/Hz. Fig. 3 illustrate the performance of 4 PSK

system over a fading channel with increases number of states (4, 8, 16, 32 and 64 states). From the simulation it is observed that, it provide diversity order and coding advantage as number of states increases with increased number of receive antennas. Table 1 depicts the comparison of frame error rate at 10 dB, 12dB, 14 and 16dB SNR (SNR Vs FER plot as shown in fig. 3) with different numbers of states.

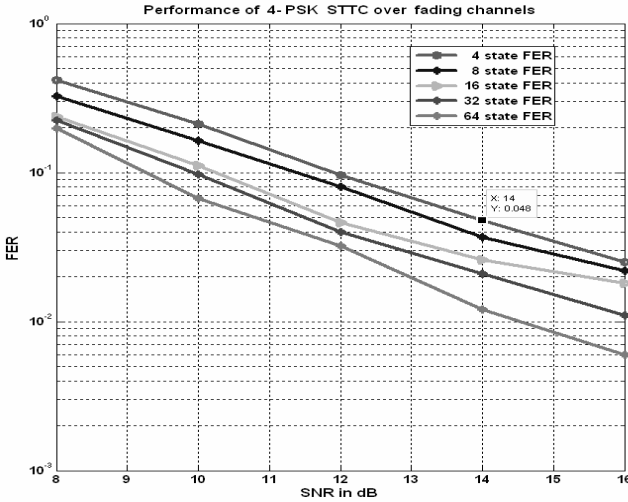


Fig. 3. Performance comparison of 4 PSK code over fading channels with $N_T=N_R=2$ antennas system

Table 1. States wise FER Comparison at 10dB, 12dB, 14dB & 16dB SNR

Sr. No.	No. of States →	4 States	8 States	16 States	32 States	64 States
1	FER at 10 dB	0.213	0.164	0.111	0.097	0.067
2	FER at 12 dB	0.096	0.08	0.046	0.04	0.032
3	FER at 14 dB	0.048	0.037	0.026	0.021	0.012
4	FER at 16 dB	0.025	0.022	0.018	0.011	0.006

6 Conclusion

In this paper, the design criteria for fading channels has been discussed and the error performance for 4 states trellis codes with two transmit and two receive antennas are investigated. It is for better performance over fading channels, codes should be designed to maximize the minimum trace of the codeword distance matrices. It provides excellent performance using 64 states encoder with coding gain and diversity. All over summary of this methodology concluded that, the system performance improved with increasing the number of states.

References

1. Foschini, G.J., Gans, M.J.: On limits of wireless communications in fading environment when using multiple antennas. *Wireless Personal Communications* 6, 311–335 (1998)
2. Tarokh, V., Seshadri, N., Calderbank, A.: Space–time codes for high data rate wireless communication performance criteria and code construction. *IEEE Trans. Info. Theory* 44, 744–765 (1998)
3. Vucetic, B.: *Space time coding*. John Wiley & Son Ltd., Chichester (2003)
4. Agrawal, D., Tarokh, V., Naguib, A., Seshadri, N.: Spac-Time coded OFDM for high data rate wireless communication over wideband channels. In: *IEEE VTC 1998* (1998)
5. Stefanov, A., Duman, T.M.: Performance bounds for Space-time trellis codes. *IEEE Transactions on Information Theory* 49(9), 2134–2140 (2003)
6. Rappaport, T.S.: *Wireless Communications: Principles and Practice*. Prentice Hall, Englewood Cliffs (1996)
7. Vucetic, B., Yuan, J.: Performance and Design of Space– Time Coding in Fading channels. *IEEE Transactions on Communications* 51(12) (December 2003)
8. Blum, R.S.: Some analytical tools for the design of space-time convolutional codes. *IEEE Trans. Commun.* 50, 1593–1599 (2002)
9. Firmanto, W., Vucetic, B.S., Yuan, J.: Space–time TCM with Improved performance on fast fading channels. *IEEE Communication Letter* 5, 154–156 (2001)
10. Haykin, S., Moher, M.: *Modern Wireless Communications*. Pearson Education, London (2009)
11. Bigleri, E., Calderbank, R., Constantinides, A., Goldsmith, A., Paulraj, A., Vincent Poor, H.: *MIMO Wireless communication*, Cambridge (2002)
12. Sanei, A.A., Ghayeb, A., Hayan, Y., Duman, T.M.: On the diversity order of space- time trellis codes with antenna selection in fast fading. *IEEE Trans. Wireless Commun.* (April 2005) (accepted for publication)
13. Hiwale, A.S., Ghatol, A.A., Bhad, S.D.: Performance Analysis Of Space Time Trellis Code with receive Antenna selection. In: *IEEE fourth International Conference on Wireless Comm. and Sensor Networks, WCSN 2008*, pp. 148–152 (December 2008)
14. Bhad, S.D., Hiwale, A.S., Ghatol, A.A.: Performance Evaluation of Space-Time Trellis Code. *IJERIA* 2(VI), 131–144 (2009) ISSN 0974-1518

Modeling Performance Evaluation of Reinforcement Learning Based Routing Algorithm for Scalable Non-cooperative Ad-hoc Environment

Shrirang Ambaji Kulkarni¹ and G. Raghavendra Rao²

¹ Research Scholar, Dept. of CSE, N.I.E, Mysore-08, India
sakulkarni@git.edu

² Professor, Dept. of CSE, N.I.E, Mysore-08, India
grrao56@gmail.com

Abstract. Scalable performance analysis of routing protocols for ad-hoc network reveals the hidden problems of routing protocols in terms of performances. Wireless nodes in ad-hoc networks may exhibit non-cooperation because of limited resources or security concerns. In this paper we model a non-cooperative scenario and evaluate the performance of a reinforcement learning based routing algorithm and compare it with ad-hoc on-demand distance vector a de facto routing standard in ad-hoc networks. Mobility models play an important role in ad-hoc network protocol simulation. In our paper we consider a realistic optimized group mobility model to aid the performance of the reinforcement learning based routing algorithm under scalable non-cooperative conditions.

Keywords: Non-cooperation, mobility models, reinforcement learning, scalability.

1 Introduction

A mobile ad hoc network exhibits special set of characteristics like dynamic topologies, variable wireless link capacity, limited energy resources and security services [1]. To achieve scalable stable performances with respect to the above characteristics is one of the most important challenges of dynamic ad hoc networks. For any form of network to be popular, it must exhibit scale economies [2]. For our investigation to meet the above challenges we consider one of the routing algorithms based on reinforcement learning known as SAMPLE [3]. The deployment of scalable applications may be for a long duration and also the data traffic may be voluminous, the limited resources of mobile nodes raise many issues. Mobile ad hoc network forms a dynamic topology of mobile nodes which lack any central authority, raising the issues of energy conservation and potential selfish behavior [4]. To aid the routing protocols in this complex interplay of mobile nodes forming dynamic topologies, we evaluate the reinforcement learning based routing algorithm SAMPLE on a Realistic Optimized Group Vehicular Mobility Model (ROPGVMM). In Section 2 we discuss some of the related work. Section 3 discusses about ROPGMM. Section 4 discusses briefly the routing protocol SAMPLE. In Section 5 we analyze the performances of routing protocols SAMPLE and AODV. Section 6 concludes the paper.

2 Related Work

J Dowling et.al [3] applied collaborative reinforcement learning algorithm SAMPLE, to enable agents to apply reinforcement learning to solve optimization problems in ad hoc networks. They considered only random waypoint mobility model with constant bit rate traffic of 4 packets per second. V Naumov and T Gross [5] in their analysis of scalability of routing methods for ad hoc networks investigated the performance of two routing protocols namely AODV and DSR. They concluded that both the protocols showed poor performance in terms of high traffic load. They considered only random waypoint mobility model. G Karakostas and E Markou [8] proposed a basic reputation-based protocol, whereby a node defines a threshold to its neighbors and drops the connection to these neighbors if they refuse to forward an amount of flow above the threshold. This served as a motivation to consider reinforcement learning algorithm SAMPLE under scalable non-cooperative environment.

3 Realistic Mobility Modeling

Our proposed model Realistic Optimized Group Vehicular Mobility Model is extended from our work [7] is as in Fig 1.

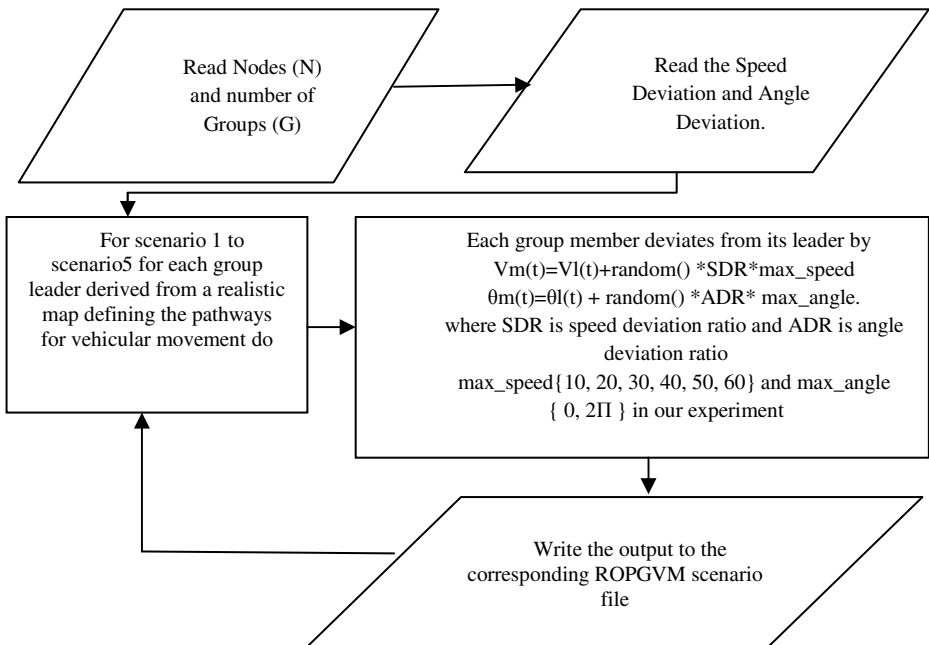


Fig. 1. Flowchart of Realistic Optimized Group Vehicular Mobility Model

4 Routing Protocols for Scalability Analysis

We consider the two routing protocols AODV and SAMPLE for investigation under scalable, non-cooperative ad-hoc environments. SAMPLE is described as follows.

4.1 SAMPLE

We consider SAMPLE [3] an example of collaborative reinforcement learning routing algorithm. The reason to consider SAMPLE for scalable and dynamic MANET operations is that it tries to optimize network operations in an adaptive manner. SAMPLE [11] routing protocol does not distinguish between separate control and data messages. This concept proves to be advantageous as only the size of packet is increased in contrast to sending of additional packets in a wireless network. In mobile ad hoc networks where congestion is a norm, some of the links may be down. AODV treats these congested links in a discrete fashion, whereas SAMPLE treats them in a continuous fashion and uses historical data in order to distinguish between a broken links from temporarily congested links.

5 Simulation Environment for Scalable Environments

NS-2 simulator ver. 2.26 from [12] has been used for scalable stable performance analysis of routing protocols like AODV and SAMPLE. The underlying MAC Protocol is defined by IEEE 802.11. Continuous bit rate (CBR) traffic sources are used. The mobility models used are the Reference Point Group Mobility (RPGM) Model and Realistic Optimized Group Vehicular Mobility (ROPGVMM) Model in a rectangular field. The field configurations are 1920 x 1920 m. The traffic generator script called cbrgen.tcl was used to generate CBR scenario of 15 sources at the rate of 4.0 kbps. The number of nodes in the simulation environment was 200 nodes. At least 5 scenarios files for RPGM [9] and 5 scenario files for ROPGVMM [7], [10] at different maximum speed of 5, 10, 15, 20, 25 sec were used for testing protocols like AODV and SAMPLE. We also extend the simulation based studies by varying the number of traffic sources i.e. 15, 20, 25, 30 and 35 at the max speed of 25 sec to test protocol performances under increasing traffic sources at high speed.

5.1 Routing Protocol Performance Metrics

The metrics [13] for evaluating the performance of AODV and SAMPLE are

Packet delivery ratio – The ratio between the numbers of packets originated by the application layer to those delivered to the final destination.

Path optimality (Average End-End Delay) – The difference between the number of hops a packet took to reach its destination and the length of the shortest path that physically existed through the network when the packet was originated.

5.2 Modeling Non-cooperative Scalable Environment

To measure the adaptability of a routing protocol, we model a non-cooperative scenario. The non-cooperation is modeled in our simulation based studies by nodes losing their transmission power below a threshold, resulting in generation of erroneous packets, which are ultimately dropped. We create this non-cooperative environment by randomly picking nodes. This results in route realignment by the routing protocols.

5.3 Non-cooperative Scalable Simulation Results

The performance of routing protocols AODV [6] and SAMPLE under scalable non-cooperative environment is illustrated in Fig 2, Fig 3 and Fig 4 and Fig 5.

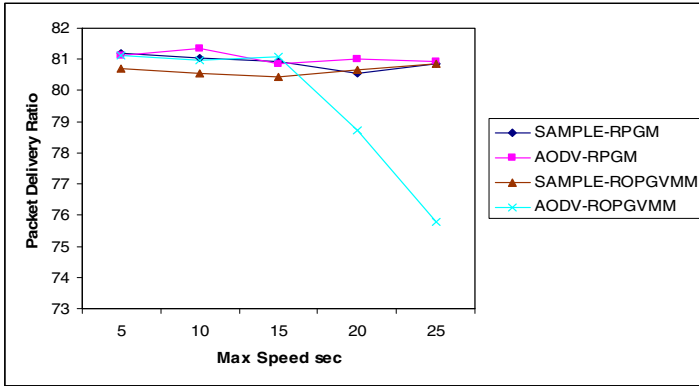


Fig. 2. Packet Delivery Fraction for AODV and SAMPLE for varying maximum speed

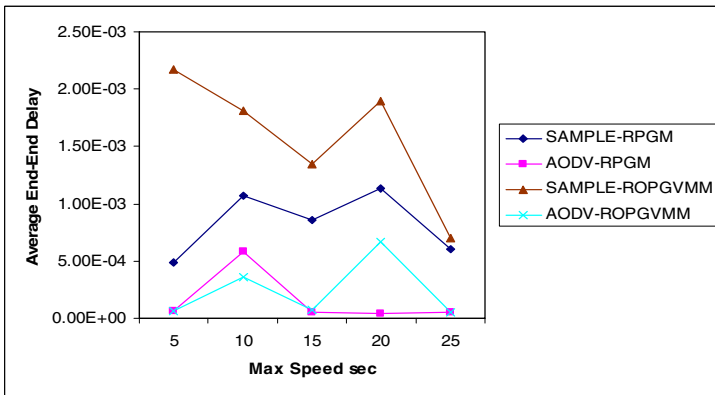


Fig. 3. Average End-End Delay for AODV and SAMPLE for varying maximum speed

In Fig. 2 we observe that the packet delivery has dropped down to the tune of 19.28%. This is due to the factor that nodes that have lost transmission energy have stopped cooperating resulting in dropping of the received packets. This ratio is more or less for AODV and SAMPLE on RPGM model. In case of ROPGVMM we observe that at high speeds AODV starts dropping more received packets and thus resulting in low packet delivery fraction. In Fig. 3 we observe that average end-end delay is poor in case of SAMPLE on both RPGM and ROPGVMM model. As the speed of nodes is increasing we observe a drop in average end-end delay of SAMPLE. This indicates the relative adaptive capabilities of SAMPLE under scalable non-cooperative environments.

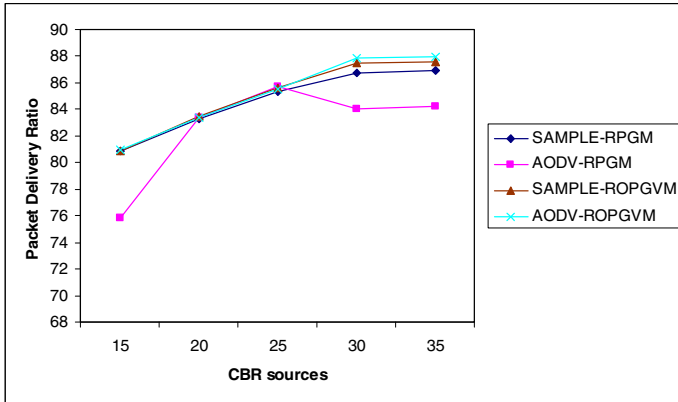


Fig. 4. Packet Delivery Fraction for AODV and SAMPLE under for varying traffic sources

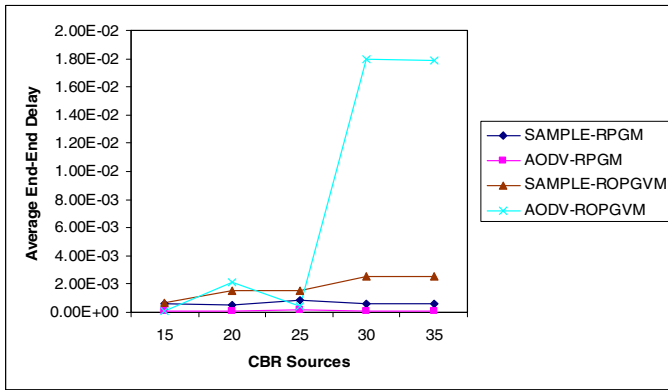


Fig. 5. Average End-End Delay for AODV and SAMPLE for varying traffic sources

In Fig 4 we observe that the packet delivery fraction under varying cbr sources in a scalable non-cooperative environment is more or less the same for AODV and SAMPLE under the RPGM model. In case of ROPGVM model AODV performs slightly poor as compared to SAMPLE which maintains a stable performance in spite of non-cooperation. SAMPLE achieves a slightly better performance as compared to AODV by 2.4% under the ROPGVM model. In Fig 5 we observe that average end-end delay under non-cooperative scalable environment and under varying traffic sources at high speed is similar for AODV and SAMPLE under the RPGM model. In case of ROPGVM model the average end-end delay of AODV protocol under high number of traffic sources is poor as compared to SAMPLE.

6 Conclusion

In our work we emphasized on performance analysis of routing protocols under scalable non cooperative scenarios. Routing protocols are critical to the design of ad

hoc networks and thus we considered two routing protocols namely AODV and SAMPLE. SAMPLE is an example of reinforcement learning algorithm and was our focus of investigation to achieve optimal performances under scalable non-cooperative environments. We observed that modeling instability in ad hoc network scenarios resulted in a low packet delivery on both models Reference Point Group Mobility Model (RPGM) and Realistic Optimized Group Vehicular Mobility Model (ROPGVM); however SAMPLE exhibited a more stable behavior both in terms of packet delivery fraction and average end-end delay on the realistic model ROPGVM as compared to AODV. Thus we can further work on a reinforcement learning approach as a routing solution for scalable realistic vehicular applications.

References

- [1] Corson, S., Marker, J.: Mobile Ad Hoc Networking (MANET): Routing Protocol Performance Issues and Evaluation Considerations, RFC 2501, pp. 3–9 (1999)
- [2] Walrand, J., Varaiya, P.: High Performance Communication Networks, 2nd edn. Morgan Kaufman, San Francisco (2005)
- [3] Dowling, J., Curran, E., Cunningham, R., Cahill, V.: Using Feedback in Collaborative Reinforcement Learning to Adaptively Optimize MANET Routing. *IEEE Transactions On Systems Man and Cybernetics. Part A: Systems and Humans* 35(3) (2005)
- [4] Urpi, A., Bonuccelli, M., Giordano, S.: Modeling cooperation in mobile ad hoc networks: a formal description of selfishness. In: Workshop: Modeling and Optimization in Mobile Ad Hoc and Wireless Networks, WiOpt3 (2003)
- [5] Naumov, V., Gross, T.: Scalability of routing methods in ad hoc networks. *Performance Evaluation Journal (Elsevier Science)* 62, 193–207 (2005)
- [6] Perkins, C., Royer, E.: Ad hoc on-demand Distance Vector Routing. In: Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications, pp. 90–100 (February 1999)
- [7] Kulkarni, S.A., Rao, G.R.: Effects of Mobility Models on Reinforcement Learning based Routing Algorithm Applied to Scalable Ad-Hoc Networks. *IJCNC* 2(6) (2010)
- [8] Karakostas, G., Markou, E.: Emergency connectivity in ad-hoc networks with selfish nodes. In: Laber, E.S., Bornstein, C., Nogueira, L.T., Faria, L. (eds.) *LATIN 2008*. LNCS, vol. 4957, pp. 350–361. Springer, Heidelberg (2008)
- [9] Hong, X., Gerla, M., Pei, G., Chiang, C.-C.: A Group Mobility Model for Ad Hoc Wireless Networks. In: *ACM/IEEE International Symposium on Modeling Analysis and Simulation of Wireless and Mobile Systems (MSWiM)*, Seattle, WA, USA (1999)
- [10] Saha, A.K., Johnson, D.B.: Modeling Mobility for Vehicular Ad Hoc Networks. In: Proceedings of the 1st ACM International Workshop on Vehicular Ad Hoc Networks, Philadelphia, PA, USA, pp. 91–92 (2004)
- [11] Curran, E.: SWARM: Cooperative Reinforcement Learning for Routing in Ad Hoc Networks, MS Thesis. Trinity College Dublin (2004)
- [12] The Network Simulator – NS-2, <http://www.isi.edu/nsam/ns/>
- [13] Broch, J., Maltz, D.A., Johnson, D.B., Hu, Y.-C., Jetcheva, J.: A performance comparison of multi-hop wireless ad hoc network routing protocols. In: *Mobile Computing and Networking (MobiCom)*, pp. 85–97 (1998)

SRPV: A Speedy Routing Protocol for VANET

Suparna DasGupta¹ and Rituparna Chaki²

¹ Department of Information Technology,
JIS College of Engineering, West Bengal, India
suparnadasguptait@gmail.com

² Department of Computer Science & Engineering,
West Bengal University of Technology, West Bengal, India
ritu.chaki@gmail.com

Abstract. Vehicular Ad hoc network is a special case of Mobile Ad hoc networks, with high nodes mobility specification and a large energy resource which could extend coverage and system lifetime. Providing reliable and efficient routing in presence of relative movement motivates the introduction of movement awareness to improve performance of existing position-based routing schemes in Vehicular Ad hoc networks. Dramatic increase in the number of vehicles equipped with computing technologies and wireless communication devices has created new application scenarios. In order to meet performance goals in VANET, several routing protocols have been formulated. Each of the existing routing protocols comes with their pros and cons. In our work we propose a new routing protocol, A Speedy Routing Protocol for Vehicular Ad hoc Network. In this routing protocol we try to reduce overhead of maintaining routing data and traverse preferred distance through shortest path.

Keywords: Mobile ad hoc networks, Routing, Mobility, Vehicular ad hoc networks.

1 Introduction

Vehicular Ad hoc Network is a special type of ad hoc networks that allow vehicles to form a self-organized network without the need for permanent infrastructure, Prerequisite to communicate between VANET are an efficient route between network nodes and this route must be adaptive in nature because of the rapidly changing topology of vehicles in motion.

A VANET is a wireless network that is formed between vehicles on an as-needed basis. Each vehicle's wireless network range may be limited to a few hundred meters so providing end-to-end communication across a larger distance requires messages to hop through several nodes. The vehicles change their location constantly. This means there is a constant demand for information on the current location and specifically for data on the surrounding traffic, routes and much more. In such networks is the non-random mobility of the vehicles; generally it is limited by roads which can be represented by digital maps. Also, the vehicle movements are limited by road rules which again may be digitally mapped.

In vehicle to vehicle communication, three broad categories of architecture are related, such as infrastructure-based, Ad hoc networks and hybrid. The infrastructure-based architecture takes advantage of the existing cellular networks. But, it has three drawbacks: high operation cost, limited bandwidth and symmetry channel allocation for uplink and downlink. Ad hoc networks do not need infrastructure, so the cost of building such network will be very low and it can even operate in the events of disasters. As a result, most research works are focus on the flexible deployment and self organizing capabilities of vehicular Ad hoc network architectures. The hybrid architecture combines these two architectures by considering vehicles as data relays between roadside base-stations. This architecture also requires the function of multi-hop communication between vehicles, which is the essential part of ad hoc network architecture.

The rest of the paper is organized in the following way. In section 2, a comparative study of some of the existing routing topologies has been carried out. We have design and describe the new routing protocol to reduce the overhead for maintaining all routing information for each vehicle, in the section no 3. Intensive performance evaluations are presented in section 4. Finally in section 5, a conclusion has been summarized.

2 Related Works

In this section we are presenting a brief comparison between some existing routing protocols. Though VANET is a subclass of MANETs, automotive ad hoc networks will behave in fundamentally different ways than predetermined models in MANET. Instead of random movement in MANETs, the movement of nodes in VANETs is constrained by the layout of roads. Normally radio range for VANETs is several hundred meters, typically between 250 and 300 meters. In a scenario when there are no radio obstacles, the nodes can communicate with others in the radio range. But in city environment, there would be radio obstacles because of buildings. Another difference is that in VANETs vehicles move with much greater speeds as compared to MANETs therefore the topology in VANETs changes much more frequently. But on the other hand the vehicles mobility can be predicted based on the speed and direction as well as the layout of roads. Therefore we have to consider all these factors for developing VANETs routing protocol.

Greedy Perimeter Stateless Routing [15] is a localized position based greedy routing protocol. It uses one hop neighbor's information to select next forwarding node. One hop neighbor's information is updated by the periodic beacon message. The algorithm has two methods for forwarding data packets: greedy forwarding and perimeter forwarding. Greedy forwarding is used whenever it possible. Perimeter forwarding is used when the greedy forwarding strategy failed.

Anchor-Based Street and Traffic Aware Routing [12] utilizes city bus route as a strategy to find routes with a high probability for delivery. It also features its own unique approach to recover from situations where its greedy strategy reaches a local optimal point. A-STAR [12] selects the farthest neighbor along a street for next hop.

Greedy Traffic Aware Routing Protocol [10] protocol makes routing decisions at street intersections, which is selected dynamically, with the selection influenced by traffic density information. For forwarding a packet each node considers the

neighboring intersections according to its street map. Based on two factors it assigns a score to each of these intersections. The first one is a measure of the traffic density between the current intersection and the potential intersection. The second feature is a measure of the distance to the destination in road length. The intersection with the highest score is selected as the next intersection for the packet.

A Movement Based Routing Algorithm [9] uses flooding for destination discovery. MORA [9] takes into account the physical location of neighbouring vehicles and their movement direction when selecting the next hop for sending/forwarding packets. The traffic overhead is large in this algorithm and it sometimes handles the case when the node goes out of the range. A Movement Based Routing Algorithm [9] is very useful in the highway scenario and it also works in the presence of non stable routes.

Connectivity Aware Routing Protocol [4] is a position-based routing protocol which also has an ability to maintain a cache of successful routes between various source and destination pairs and forwarding the data packet along the found path. It provides better scalability, performance and robustness against frequent topological changes. The connectivity Aware Routing Protocol [4] is suitable for both dense and sparse networks and it also takes care of high node mobility. In this routing protocol data packet delays are comparatively less and data delivery ratio also improved and it provides better network throughput. Positions needs to be known and use of GPS from navigation systems are also important.

Adaptive Connectivity Aware Routing Protocol [6] is a protocol which adaptively selects an optimal route with the best network transmission quality based on the statistical and real time density data that are gathered through an on-the-fly density collection process. ACAR [6] calculates a path based on the statistical density data and then the path information is included into packet headers and the packets are transmitted along the selected path. Adaptive Connectivity Aware Routing Protocol [6] works well even if accurate statistical data of road density is not available. Data delivery ratio is increased and transmission delay is decreased in the ACAR [6] protocol. Better throughput performance is also achieved in this protocol. Though route length has to be calculated, more overhead also considered in this technique.

Position Based Routing Protocol [8] is obtaining location and velocity information of vehicles on the route to the gateway and the prediction algorithm that uses this information to predict when the route will break. An additional consideration is whether to use routes through vehicles that are not moving in the same direction as the vehicle requiring the route, i.e., the source node. PBR [8] is specially tailored to the mobile gateway scenario and takes advantage of the predictable mobility pattern of vehicles on highways. PBR [8] protocol offers from Ad hoc On Demand Distance Vector [14] and DSR [13] in the way that it proactively creates new routes before they break.

Directional Greedy Routing Protocol [5] uses location speed and direction of motion neighbours to select most appropriate next forwarding nodes. It uses two forwarding strategies greedy and perimeter. It uses speed and direction of its neighbours. DGRP [5] leads to a better choice to select next forwarding node, i.e. forwarding node determines node which is more appropriate at the time of forwarding. In this protocol each node can get information about nodes coming into

its communication range from its neighbours. Directional Greedy Routing Protocol [5] does not suffer from any looping problem unlike ARP [11] because it does not use compass routing. In DGRP [5], each node has to compute the speed and direction of motion and they have to predict the position of all its neighbours. This protocol does not send beacon packet on need basis, i.e. which is proportional to speed, which leads to increase in overhead and complexity. DGRP [5] selects better next hop node than MOPR [7] because MOPR [7] selects only those nodes for forwarding which are going to be communication range for next one second.

In the next section we are going to propose a new routing protocol and try to reduce the problems of previously discussed routing protocols.

3 Proposed New Routing Protocol

The study of the state of art reveals that most of them suffer from large overhead of maintaining large amount of routing data and maintaining data about high speed vehicle. To overcome these drawbacks, a new routing protocol is proposed. In this routing protocol we use the co-ordinate system for distance calculation between two vehicles. This gives a clear idea of destination vehicle's location. We select the vehicle that has less mobility and preferred distance from sender vehicle.

The activity of the algorithm is described below:

Table 1. Data Dictionary

Variable	Description
req_{msg}	request message
rng_{max}	maximum communication range
TTL_1	Time to Live
cnt	counter
ngh_{vhid}	neighbor vehicle id
$dest_{id}$	Variable
$destination_{id}$	target destination vehicle id
s_{pd}	speed of vehicle
c_r	communication range
TTL_2	Time to Live
p_{ack}	positive acknowledgement
n_{ack}	negative acknowledgement

Table 1. (continued)

msg_{id}	message id
$home_{id}$	receiver vehicle id
$recV_{msgid}$	received message id
$prefdist_{min}$	preferred minimum distance
$prefdist_{max}$	preferred maximum distance
src_{vhid}	source vehicle id
cnt_{msg}	message count
$vhl[i]$	vehicle id

A. Operation Performed at Sender Side

At first, the sender vehicle broadcasts a req_{msg} to all the neighbor vehicles, within its communication range i.e. rng_{max} . At that very instant, the sender vehicle starts a timer that ticks till a certain period of time i.e. TTL_1 . Within this time if no acknowledgement is received from any neighbor vehicles then the req_{msg} is broadcast again. If within this time period, any acknowledgement is received from any of the neighbors then a counter, which was previously set to the value '0', is incremented its value by '1'. For every acknowledgement is received, and the acknowledgements are stored sequentially as per their arrival. Then the acknowledgements are taken one by one and the ngh_{vhid} is extracted. The ngh_{vhid} is then stored in a variable called $dest_{id}$. The content of $dest_{id}$ is then compared with the $destination_{id}$ to find out whether that particular vehicle is the destination or not.

If the contents match then it implies that the vehicle is the destination and the message is delivered to the destination. A timer is started at that very instant and the sender vehicle waits for a certain fixed period of time, i.e. TTL_2 for the acknowledgement to be received from the destination neighbor vehicle. If the acknowledgement is received within the specified time i.e. TTL_2 , the sender vehicle stops further processing of the message and waits for a new task. In case no acknowledgement is received the sending process is continued again.

Now, if the neighbor vehicle is not the final destination, then the distance of the neighbor vehicle from the sender vehicle is measured and stored [1][2]. If the distance lies within the preferred range, which is predefined, then the speed (s_{pd}) of the neighbor vehicle is calculated. If the speed (s_{pd}) lies within a pre-defined preferred range, then the message is prepared to be sent and send to the neighbor vehicle along within its communication range (c_r).

Again, a timer is started and the sender vehicle waits for the acknowledgement to be received from the neighbor vehicle. If no acknowledgement is received within the specified time i.e. TTL_2 , the message is delivered again. If p_{ack} is received from the neighbor vehicle and the entire process is completed.

If n_{ack} is received then rng_{max} of the vehicle is extracted from the acknowledgement and the c_r is reset to a value less than the rng_{max} of the vehicle and the message is resend along with the new communication range ($c_r \leq rng_{max}$).

B. Operation Performed at Receiver End

• On Receiving Request Message

When a req_{msg} is received by the receiver vehicle, it extracts the ngh_{vhlid} from it and sends an acknowledgement to that sender vehicle.

• On Receiving Message

When the receiver vehicle first receives a message, along with the msg_{id} , ngh_{vhlid} and the destination $_{id}$. Then the receiver vehicle checks the destination $_{id}$ with its own id, i.e. $home_{id}$. If the $home_{id}$ matches with the destination $_{id}$, then another check is to be conducted to find out whether the msg_{id} matches with any $recv_{msgid}$. If it matches then the new message is discarded and an acknowledgement is sent to the sender. If the new message id and the previous message id are different, then the message is accepted and displayed and the cnt gets incremented by '1'.

Now, if the $home_{id}$ and the destination $_{id}$ do not match then, c_r is extracted from the message. A check is conducted whether the rng_{max} of the receiver vehicle is less than the c_r as sent by the sender vehicle. If rng_{max} is less than c_r then the message is discarded and a n_{ack} is sent to the sender vehicle along with the value of rng_{max} . Or else, a p_{ack} is sent and the receiver vehicle then acts like the sender vehicle and then the full function of sender vehicle is followed.

☑ sender ()

Step 1: after the compilation of the message, the message is made ready to be sent.

Step 2: $rng = 0$

Step 3: if ($(rng < rng_{max}) \ \&\& \ (receiver \ found)$)

Step I: broadcast req_{msg}

Step II: $rng ++$

Step III: GOTO Step 3

Step 4: start TIMER

Step 5: $\Delta t = 0$

Step 6: if ($\Delta t < TTL_1$)

Step I: wait for ACK

Step II: $cnt = 0, i = 0$

Step III: if (ack[i] received)

Step a: $cnt ++$

Step b: $i ++$

Step c: GOTO Step III

Step 7: if ($(\Delta t = TTL_1) \ \&\& \ (ack \ [i] \ received = \ NULL)$)

Step I: GOTO Step 2

Step 8: $i = 0$

Step 9: if ($i \leq cnt$)

Step I: extract ngh_{vhlid} from ack[i]

Step II: $dest_{id} = ngh_{vhlid}$

Step III: $i ++$

Step IV: GOTO Step 9

Step 10: if ($dest_{id} = destination_{id}$)

- Step I: unicast MSG to destination
- Step II: start TIMER
- Step III: $\Delta t = 0$
- Step IV: if ($\Delta t < TTL_2$)
 - Step a: wait for ack
 - Step b: if (ack [i] received == NULL)
 - Step i: GOTO Step 2
 - Else exit
 - Step c: $\Delta t ++$
 - Step d: GOTO Step IV

Step 11: else

- Step I: $i = 0$
- Step II: if ($i \leq cnt$)
 - Step a: measure Δd for $vh[i]$ and store
 - Step b: if ($(\Delta d \geq \text{prefdist}_{min}) \ \&\& \ (\Delta d \leq \text{prefdist}_{max})$)
 - Step i: check s_{pd}
 - Step ii: if ($s_{pd} < s_{pdmax}$)
 - Step A: set c_r
 - Step B: prepare msg including c_r
 - Step C: unicast msg to $vh[i]$
 - Step D: start TIMER
 - Step E: $\Delta t = 0$
 - Step F: if ($\Delta t \leq TTL_1$)
 - Step A1: wait for ACK
 - Step A2: $\Delta t ++$
 - Step A3: GOTO Step F
 - Step G: if (ack received == NULL)
 - Step A1: GOTO Step C
 - Step H: else
 - Step A1: if (p_{ack} received)
 - Step B1: exit (0)
 - Step A2: else
 - Step B2: if (n_{ack} received)
 - Step C1: extract rng_{max}
 - Step C2: set $cr = rng_{max} - 1$
 - Step C3: prepare msg with c_r
 - Step C4: unicast msg to $vh[i]$

Step 12: end.

☑ receiver ()

On Receiving Request Message

Step 1: extract src_{vhid} from req_{msg}

Step 2: send ack to src_{vhid}

On Receiving Message

```

Step1: extract msgid, srcvhlid, destinationid
Step 2: if (destinationid == homeid)
    Step I: i = 0
    Step II: if (i < cntmsg)
        Step a: if (msgid == recvmsgid [i])
            Step i: send ack
            Step ii: discard the msg
        Step b: i ++
        Step c: GOTO Step II
    Step III: else
        Step a: cntmsg ++
        Step b: display msg
Step 3: else
    Step I: extract cr from msg
    Step II: if (rngmax <= cr)
        Step a: discard msg
        Step b: send nack along with rngmax
    Step III: else
        Step a: send pack
        Step b: sender ( )
Step 4: end.

```

By using the above discussed algorithm a vehicle can communicate with any another vehicles within that network.

4 Performance Analysis

The simulation model consists of a network model that has a number of vehicular nodes, which represents the entire network to be simulated. The main objective of this algorithm is to reduce the overhead of maintaining route information for vehicular ad hoc network. For this reason, in the place of maintaining information about all vehicular nodes we have maintained information only about the next neighbor vehicle node.

The graph in fig 1 shows the time required for finding route from source to destination for a different number of vehicular nodes. From that it is clearly seen that when the number of nodes in the network are increased then the required time for route finding also increased.

In fig 2, there is a comparison between number of nodes and hop-count for finding route from source to destination. Here we have also observed that if the number of nodes increase, initially hop count also increase. But after a certain value, hop count fixed in a certain range.

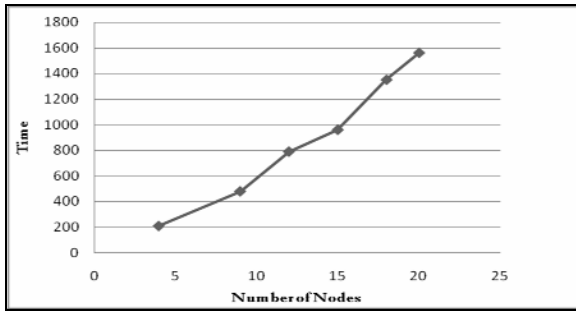


Fig. 1. Number of Nodes vs. Hop Count

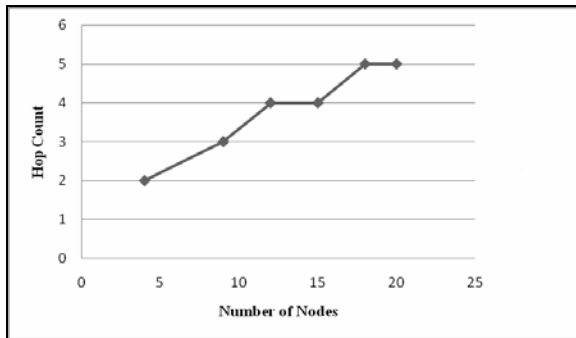


Fig. 2. Number of Nodes vs. Hop Count

5 Conclusions

In this paper a new routing protocol has been proposed. In which we try to reduce the overhead of maintaining a large amount of data by selecting vehicles with less speed. So a large amount of overhead has been reduced by minimizing the up-ration of routing tables. Packets are routed to the neighbour node at largest distance from the sender, in an attempt to cover maximum distance in one hop. The source vehicle stores information only about neighbour vehicles nodes. The performance is being studied now, and the initial results are included, which shows that the algorithm behaves reasonable well with a sparse network. Research is now on to verify its performance in a dense network and arrive at a comparative analysis of performance with recent algorithms.

References

1. DasGupta, S., Chaki, R.: AMOBIRROUTE: An Advanced Mobility Based Ad Hoc Routing protocol for Mobile Ad Hoc Networks. In: IEEE International Conference on Networks & Communications, NetCom 2009 (2009)
2. Saha, S., DasGupta, S., Chaki, R.: MOADRP: Mobile Ad hoc Network Routing Protocol. In: IEEE International Conference on Wireless Communication and Sensor Networks, WCSN 2009 (2009)

3. Saha, S., DasGupta, S., Chaki, R.: A Survey of Prediction-Based Routing Protocols for Vehicular Ad hoc Network. In: 12th International Conference on Information Technology, ICIT 2009 (2009)
4. Yang, Q., Lim, A., Agrawal, P.: Connectivity Aware Routing in Vehicular Networks. In: IEEE Communications Society, WCNC 2008 (2008)
5. Kumar, R., Rao, S.V.: Directional Greedy Routing Protocol (DGRP) in mobile Ad hoc network. In: International Conference on Information Technology, ICIT 2008 (2008)
6. Yang, Q., Lim, A., Li, S., Fang, J., Agrawal, P.: ACAR: Adaptive Connectivity Aware Routing Protocol for Vehicular Ad Hoc Networks. In: IEEE Conference (2008)
7. Menouar, H., Lenardi, M., Filali, F.: Movement Prediction-based Routing (MOPR) Concept for Position-based Routing in Vehicular Networks. In: Eurocom (2007)
8. Nambodiri, V., Gao, L.: Prediction-Based Routing for Vehicular Ad Hoc Networks. IEEE Transaction on Vehicular Technology 56(4) (July 2007)
9. Granelli, F., Boato, G., Kliazovich, D.: MORA: a Movement- Based Routing Algorithm for Vehicle Ad Hoc Networks. In: IEEE Workshop on Automotive Networking and Applications (AutoNet 2006), San Francisco (December 2006)
10. Jerbi, M.: Greedy Traffic Aware Routing Protocol. In: International Conference on Mobile Computing & Networking, Los Angeles (2006)
11. Giruka, V.C., Singhal, M.: Angular Routing Protocol for Mobile Ad hoc Networks. In: 25th IEEE International Conference on Distributed Computing Systems Workshops, ICDCSW (2005)
12. Liu, G., Lee, B.S., Seet, B.C., Foh, C.H., Wong, K.J., Lee, K.K.: A routing strategy for metropolis vehicular communications. In: Kahng, H.-K., Goto, S. (eds.) ICOIN 2004. LNCS, vol. 3090, pp. 134–143. Springer, Heidelberg (2004)
13. Johnson, D., Maltz, D., Jetcheva, J.: The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks. Internet Draft, draft-ietf-manet-dsr-07.txt (work in progress) (2002)
14. Das, S., Perkins, C., Royer, E.: Ad Hoc On Demand Distance Vector (AODV) Routing. Internet Draft, draft-ietf-manet-aodv-11.txt (2002)
15. Karp, B., Kung, H.T.: GPSR: Greedy Perimeter Stateless Routing for Wireless Networks. In: Proc. of ACM/IEEE International Conference on Mobile Computing & Networking 2000 (August 2002)

Simultaneous Multiple Link/Node Failure Handling for Different Service-Paths in MPLS Networks

Shah Rinku¹ and Chatterjee Madhumita²

¹ Vidyalkar Institute of Technology, Wadala(E), Mumbai, India

² Ramrao Adik Institute of Technology, Nerul, Mumbai, India
rinku.shah@vit.edu.in, madhumita@rait.ac.in

Abstract. Failure handling and recovery under real-time conditions is very crucial. MPLS fast reroute was designed to meet the needs of real-time applications, such as voice over IP. Few of existing solutions handle failure based on single-link failure assumption. Few talk about multiple link failure handling but for a single service path. The proposed solution, MLNF (Simultaneous Multiple Link/Node Failure handling for different service-paths) presents a dynamic solution, which handles simultaneous multiple-link failures as well as node failure that may occur on different service paths. An algorithm and metric is proposed for optimized and prioritized backup path computation. A novel dynamic metric calculation for service path computation is also proposed. Our proposed algorithms are simulated using ns2. The simulation results show that MLNF provides an improvement over existing LDP Reroute method.

Keywords: MPLS, failure handling, Qos, LSP, Traffic engineering, LDP.

1 Introduction

Today the Internet world is more inclined towards MPLS networks since they provide fast switching than traditional IP networks. Because of this network resilience in MPLS/GMPLS networks is currently receiving considerable attention in research and standardization communities [1], since link and/or node failures in MPLS/GMPLS networks may incur tremendous amount of significant data loss of critical business applications and services, if the failures are not recovered properly. Resilience is the capability of recovering from network component failures. Multiple simultaneous failures can often occur in a large-scale MPLS/GMPLS network infrastructure [3].

IETF is currently actively working on the standardization of the MPLS recovery mechanism [5], [6], which can be roughly classified into the following two techniques: protection and rerouting.

We propose a metric for balanced load, optimized service-path computation. We propose a novel approach for dynamic, prioritized and optimized backup path computation in advance for simultaneous failure of different service path. The proposed algorithms also look into the problem of TE and provide load balancing over each edge.

The paper is organized as follows: Section 2 describes our proposed algorithm MLNF. Section 3 discusses the simulation and results. Section 4 concludes our work and Section 5 talks about the future scope.

2 Proposed Solution: Multiple Link/Node Failure Handling (MLNF)

MLNF is a centralized, dynamic solution to handle multiple link failure and node failure. MLNF also provides load balancing by proposing a novel metric for service path computation. MLNF also proposes an algorithm for backup-path computation.

When an LSP traffic request is received by the central controller, service path is computed and backup-path is also computed in advance according to the priority of the LSP traffic request.

If failure occurs, traffic could instantaneously switch over to the computed backup path and when the service-path recovers from failure, traffic switches back to the service-path.

2.1 Service Path Computation

For every LSP request, we have parameters like requested traffic bandwidth for service path (REQ_s) and priority (P) of the LSP request. The priority here specifies the level of protection requested by LSP request if failure occurs. Reservation $\% = P * 25\%$

For service path computation, we would use Dijkstra's algorithm but we provide a new metric for cost of an edge in the network as follows.

$$C_{i,j} = (((A[i,j] - R[i,j]) / A[i,j]) + ql_j) * W[i,j] \quad (1)$$

where ql_j : represents the queue length of the downstream router, $R[i,j]$ is the reserved bandwidth on link (i,j) , $A[i,j]$ is available bandwidth on link (i,j) and $W[i,j]$ is the weight of link (i,j) which could be any parameter like cost or distance. We have considered the queue length to reduce delay along the path and $((A[i,j] - R[i,j]) / A[i,j])$ factor to restrict too much of reservation on the same edge.

These computations are done dynamically every time a new service path or backup path is computed.

2.2 Computation of Possible Backup Paths

2.2.1 Data Structures Used:

Available bandwidth for edge (i,j) : $A[i,j]$

Requested traffic bandwidth for service path P_s : REQ_s

$\%$ Reservation bandwidth for backup path: $RSV\%$ which is computed as $P * 25\%$

ServPath[y]: List of vertices in service path excluding ingress and egress

AllPaths[x]: Structure holding all paths for source (ingress) to destination (egress).

CurPath[p]: List of vertices representing current path for source (ingress) to destination (egress).

Vertex[s]: Source (ingress), Vertex[d]: Destination (egress), [c]: Current vertex

Initially [c] is the source (ingress)

2.2.2 Algorithm for Computation of Possible Backup-Path

```

Backup_path_calc (vertex [c])
begin
  If (P == 0)
    Return
  else
    begin
    Add [c] to tail-end of CurPath[p]
    For each vertex [v] adjacent to [c]
    begin
      If [v] is equal to [d]
      begin
        Add [v] to CurPath[p]
        Save CurPath[p] in AllPaths[x]
      end
      else
      If (([v] is not in ServPath[y]) AND ([v] is not in
      CurPath[p]))
      begin
        Add [v] to CurPath[p]
        Backup_path_calc ([v])
      end
    end
    end
  end
  Return
end

```

Backup_path_calc (vertex [c]) computes all possible disjoint backup-paths from ingress to egress of a service path. That is the backup path will have no vertex or edge in common with that of service-path. This ensures that even if one edge or vertex of service-path fails our backup-path will be able to handle the failure.

The output is stored in AllPaths [] which can be concurrently accessed by Resilience_requirement_check () algorithm hence reducing computation time.

2.2.3 To Check Whether the Backup Path Satisfies the Resilience Requirements

Resilience_requirement_check () algorithm checks whether the backup-path satisfies the protection requirement requested by service-path. If for any edge in the path $((REQ_s * RSV \%) < A [i, j])$, path is discarded from AllPaths [] .

2.2.4 Choice of Backup-Path for a Service-Path

Backup_path_cost for k^{th} path (BPC_k) is as follows:

$$BPC_k = \left\{ \sum_{n=1}^N q_n + \sum_{e=1}^E BPV_{i,j} \right\} \sum_{e=1}^E W[i, j] \quad (2)$$

where N is total number of nodes in the backup path ,E is total number of edges of backup path, $BPV_{i,j}$ is Backup path vector value for edge i, j and $W [i, j]$ is weight of edge i, j. Choose the backup path stored in AllPaths[x] which has minimum BPC_k .

Backup path vector (BPV) is a $k \times k$ matrix which tells us about a particular edge is used as a part of backup path by how many backup paths. The lesser the number of

backup paths using certain edge the better. This is because the amount of reserved bandwidth i.e. currently unutilized bandwidth is high.

3 Simulation Results

MLNF is simulated on NS 2.27, MNSv1.0 installed on Windows XP platform. MNS includes the LDP patch which provides default traffic engineering. MLNF uses LDP as the signaling protocol but the service LSP and backup LSP is setup as an explicit route using MLNF.

3.1 Performance Metrics

We have used the following performance metrics for our analysis.

- **Packet Dropping Ratio** – This metric specifies the ratio of total number of packets dropped to total number of packets sent.
- **Cumulative Queue Length at each node** – This metric monitors the congestion level at each node.
- **Cumulative Queue Length at each Link** – This metric is used to measure the load balancing over the links of the topology.

3.2 Example Topology

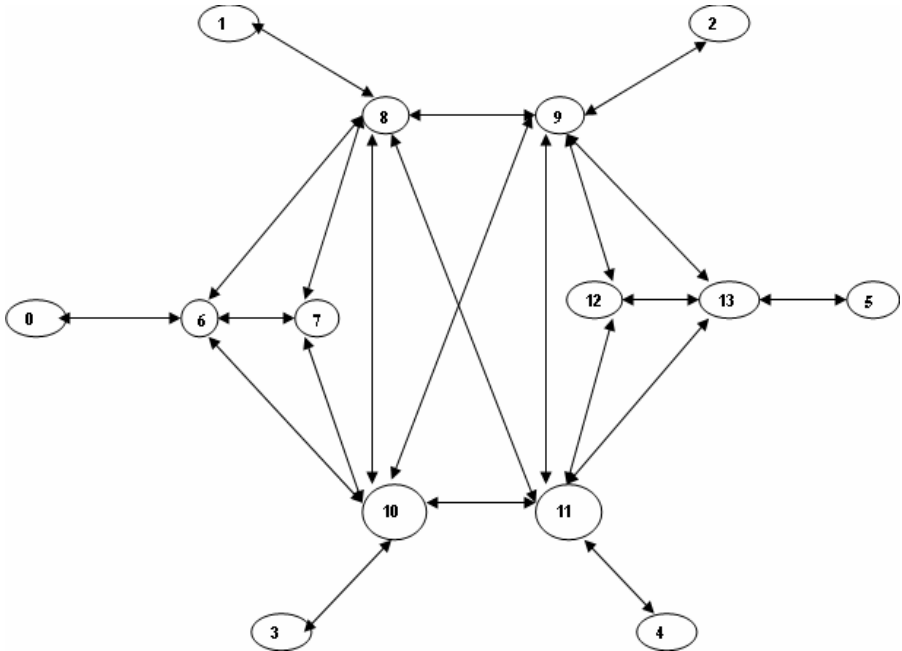


Fig. 1. Example Topology

*Weight matrix $W [i, j]$ is unit cost and Initially Reserved bandwidth $R [i, j]$ for MPLS nodes is 0

Table 1. Available bandwidth $A [i, j]$ for MPLS nodes before traffic requests

	6	7	8	9	10	11	12	13
6	0	10	10	0	10	0	0	0
7	10	0	6	0	10	0	0	0
8	10	6	0	10	6	10	0	0
9	0	0	10	0	10	6	6	10
10	10	10	6	10	0	10	0	0
11	0	0	10	6	10	0	10	10
12	0	0	0	4	0	10	0	10
13	0	0	0	10	0	10	10	0

Table 2. Traffic requests for Topology1

Traffic Req No.	LSP request	Ingress-Egress	Requested bandwidth (in Mbps)	Priority	Service path	Chosen backup path (BPC _{min})
1	0-5	6-13	6.4	4	6-8-9-13	6-10-11-13
2	3-2	10-9	3.2	4	10-6-8-9	10-9
3	1-2	8-9	1.6	2	8-6-10-9	8-11-9
4	1-5	8-13	1.6	2	8-11-13	8-10-9-12-13
5	4-1	11-8	3.2	4	11-8	11-10-8

The results are computed using the equations and algorithms mentioned in Section 3.

Table 3. Failure cases for Topology1

Case	Failed LSP's
No failure	-
Link 8-9,8-11 failure	0-5,3-2,1-5,4-1
Node 6 failure	0-5,3-2,1-2
Node 8 failure	0-5,3-2,1-2,1-5

3.3 Result Analysis

LDP Reroute ■ MLNF ■

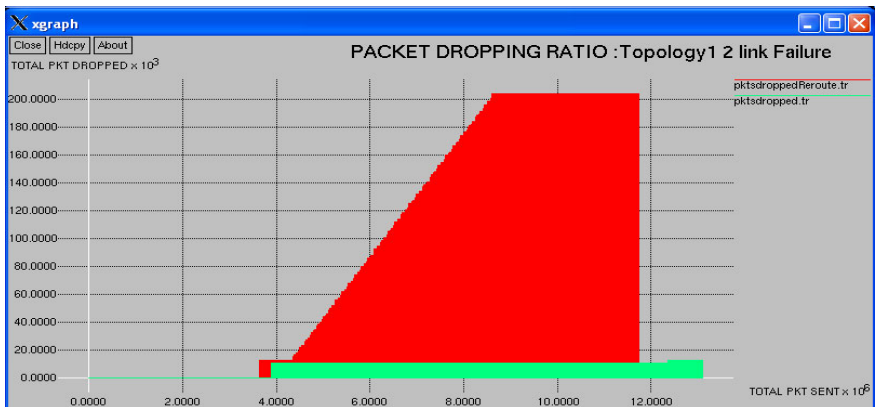


Fig. 2. Packet Dropping Ratio: Topology 1 Two Link failure

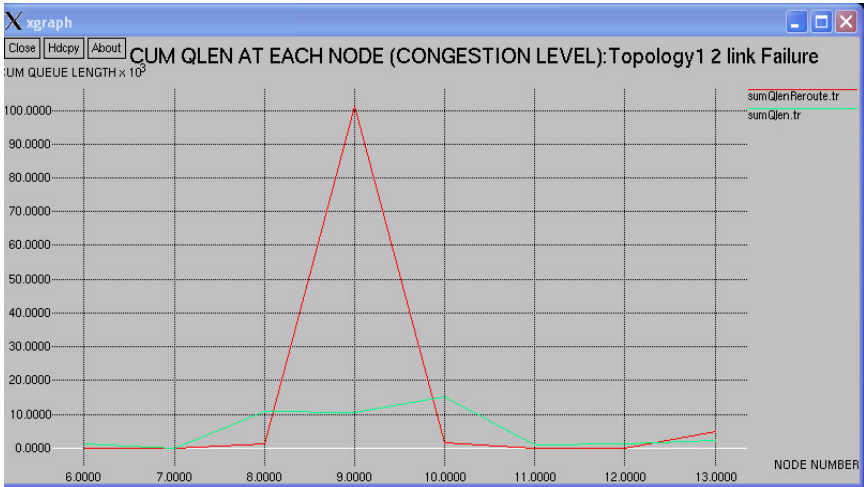


Fig. 3. Congestion Level: Topology 1 Two Link failure

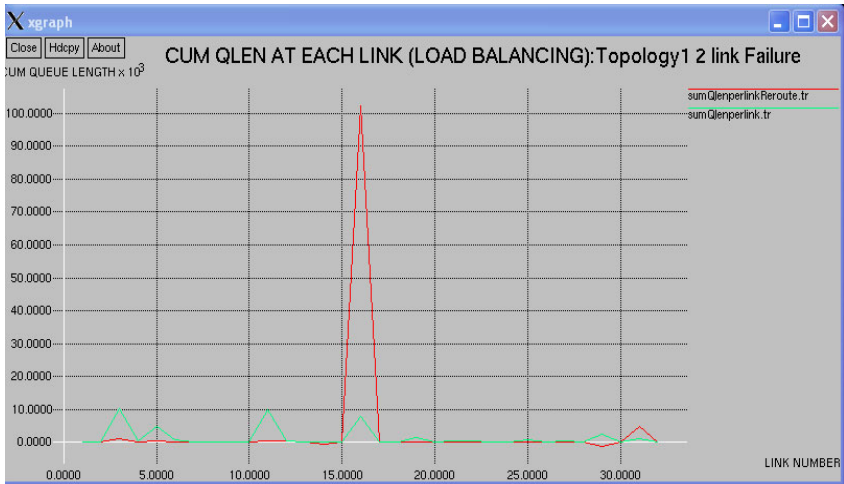


Fig. 4. Load balancing: Topology 1 Two Link failure

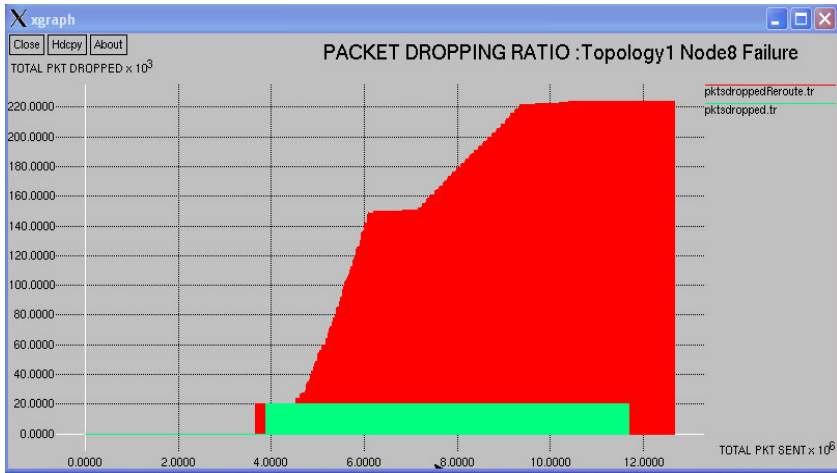


Fig. 5. Packet Dropping Ratio: Topology 1 Node 8 failure

4 Conclusion

MLNF proposes a new dynamic metric for service-path computation. MLNF also provides a dynamic and prioritized backup-path computation algorithm which handles simultaneous multiple link failures as well as node failures on different service-paths whereas all other existing solutions are based on assumption of single LSP or single link failure.

MLNF is successful in balancing the load on all links in most cases and provides lower packet dropping ratio in all cases. MLNF aims at utilizing all the links instead of always choosing the shortest path as in Traditional IGP.

MLNF does not always provide good end-to-end delay in case there is no failure where LDP_Reroute works better.

5 Future Work

We have proposed a centralized algorithm “MLNF”. Our future work would be in converting into a partial distributed approach. Partial distributed approach because completely distributed approach may use up unnecessary bandwidth.

References

1. Lang, J.P., Drake, J.: Mesh Network Resiliency using GMPLS. Proc. IEEE 90(9), 1559–1564 (2002)
2. Iraschko, R.R., Grover, W.D.: A Highly Efficient Path- Restoration Protocol for Management of Optical Network Transport Integrity. IEEE J. Selected Areas in Comm. 18(5), 779–794 (2000)

3. Wang, J., Sahasrabudde, L., Mukherjee, B.: Path versus Subpath versus Link Restoration for Fault Management in IP-over-WDM Network: Performance Comparisons Using GMPLS Control Signaling. *IEEE Comm. Magazine* 40(11), 80–87 (2002)
4. Markopoulou, Iannaccone, G., Bhattacharya, S., Chuah, C.-N., Diot, C.: Characterization of Failures in an IP Backbone. In: *Proc. IEEE INFOCOM 2004*, vol. 4(7-11), pp. 2307–2317 (March 2004)
5. Papadimitriou, D., Mannie, E.: Analysis of Generalized Multi- Protocol Label Switching (GMPLS)-Based Recovery Mechanisms (Including Protection and Restoration), RFC 4428 (March 2006)
6. Lang, J., Rajagopalan, B., Papadimitriou, D.: Generalized Multi- Protocol Label Switching (GMPLS) Recovery Functional Specification, RFC 4426 (March 2006)
7. Park, J.T., Nah, J.W., Lee, W.H.: Dynamic Path Management with Resilience Constraints under Multiple Link Failures in MPLS/GMPLS Networks. *IEEE Transactions On Dependable And Secure Computing* 5(3) (July-September 2008)
8. Wang, D., Li, G.: Efficient Distributed Bandwidth Management for MPLS Fast Reroute. *IEEE/ACM Transactions on Networking* 16(2) (April 2008)
9. Managing Your Migration TO MPLS by Network Physics, Inc. (July 2004)

Energy-Efficient Multilevel Clustering in Heterogeneous Wireless Sensor Networks

Vivek Katiyar, Narottam Chand, and Surender Soni

Department of Computer Science and Engineering
National Institute of Technology Hamirpur
Hamirpur (H.P.), India

{vivek.kat,nar.chand,surender.soni}@gmail.com

Abstract. Researchers generally believe that nodes in wireless sensor networks (WSNs) are homogeneous, but some sensor nodes of higher energy can be used to prolong the lifetime and reliability of WSNs. This gives birth to the concept of Heterogeneous Wireless Sensor Networks (HWSNs). Clustering is an important technique to prolong the lifetime of WSN and to reduce energy consumption as well, by topology management and routing. HWSNs are very popular in real deployments [1], and have a large area of coverage. In such scenarios, for better connectivity, the need of multilevel clustering protocols arises. In this paper we propose an energy efficient protocol called heterogeneous multilevel clustering and aggregation (HMCA) for HWSNs. We simulate and compare HMCA with existing multilevel clustering protocol EEMC [2] for homogeneous WSN. Simulation results demonstrate that our proposed protocol performs better.

Keywords: data aggregation, energy efficiency, heterogeneity, multilevel clustering.

1 Introduction

In recent years, with the advances in the technology of micro-electromechanical system (MEMS) and developments in wireless communications, wireless sensor networks (WSNs) have gained worldwide attention. WSNs consist of small nodes (sensors) having capabilities like sensing, computation, and communications. These sensors gather data by sensing the surroundings, aggregate this data to form useful information and transmit it to the base station or the neighboring node. Sensor nodes are limited in power, memory and computational capacity. So they may be short lived. One of the ways to prolong the lifetime of WSN is to insert a percentage of sensor nodes equipped with additional energy resources i.e. making the WSN heterogeneous in terms of energy. Many existing schemes for heterogeneous wireless sensor networks (HWSNs), like SEP [3], EEHC [4], DEEC [5], etc., demonstrate that HWSNs survive for a longer time, as compared to homogeneous WSNs.

With the advances in MEMS, VLSI technology and wireless communication technology, it has become possible to form a large scale WSN with small sensor nodes scattered across some environment. These sensor nodes need to be in touch with far-away base station. Clustering is an important technique to prolong the

lifetime of WSNs and to reduce energy consumption as well, by topology management and routing. Sometimes, direct communication may not be possible with the base station due to certain limitations. This leads to the need of multilevel clustering algorithms for large WSNs.

An Energy Efficient Multilevel Clustering Algorithm (EEMC) to minimum energy consumption in homogeneous WSNs is proposed in [2]. EEMC also covers the cluster head election scheme. Like other clustering algorithms [3, 4], EEMC also works in two phases: cluster set-up phase and data transmission phase.

The scheme assumes that the sensor network is homogeneous. In our work, we apply the concept of multilevel clustering to HWSNs. Simulation results are compared with multilevel clustering scheme EEMC for homogeneous WSNs.

Rest of the paper is organized as follows. Section 2 presents the assumptions and heterogeneous model for WSN. Our proposed algorithm, HMCA, is described in section 3. Section 4 presents the simulation environment and results. We conclude the paper in section 5.

2 Preliminaries

2.1 Assumptions

This paper considers a HWSN deployed for real life applications. The following assumptions are made about the sensor nodes and the network model:

1. The base station (i.e. sink node) is located inside the sensing field.
2. Nodes in the sensor field are heterogeneous in terms of energy.
3. Communication within the square area is not subjected to multipath fading.
4. The communication channel is symmetric.
5. Data gathered can be aggregated into single packet by cluster heads (CH).

2.2 Heterogeneous Model of WSNs

Here, we present a paradigm of HWSN and discuss the impact of heterogeneous resources described in [7]. There are three common types of resource heterogeneity in sensor nodes: computational heterogeneity, link heterogeneity and energy heterogeneity.

Computational heterogeneity means that the heterogeneous node has a more powerful microprocessor and more memory, than the normal node. Link heterogeneity means that the heterogeneous node has more bandwidth and a longer distance network transceiver than the normal node. Energy heterogeneity means that the heterogeneous node is line powered, or its battery is replaceable. Among above three types of resource heterogeneity, the most important heterogeneity is the energy heterogeneity because both computational heterogeneity and link heterogeneity will consume more energy resources.

If some heterogeneous nodes are placed in the sensor network, the average energy consumption will be lower in forwarding a packet from the normal nodes to the sink, and hence the network lifetime will increase. Reliability of data transmission is also improved with the help of heterogeneous nodes.

3 Proposed Algorithm

Most of the proposed multilevel clustering schemes consider WSN to be homogeneous. EEMC and PAMC [6], both algorithms do not consider the heterogeneity of nodes in terms of their initial energy. HMCA deals with the adaptation of the CH election process considering heterogeneous nodes in a multilevel clustering scheme.

The model proposed here for heterogeneous WSN is same as [5] that considers two types of nodes: normal nodes and advance nodes. It is assumed that out of total N nodes in the network, there are $(1 - \theta) * N$ normal nodes and $\theta * N$ advance nodes. Advance nodes have γ times more energy than the normal nodes. The initial energies of normal nodes and advance nodes are E_{ini} and $\gamma * E_{ini}$ respectively. Hence total energy of network

$$\begin{aligned} E_{total} &= N * (1 - \theta) * E_{ini} + N * \theta * E_{ini} * \gamma \\ &= N * E_{ini}(1 - \theta + \theta * \gamma) \\ &= N * E_{ini} + N * E_{ini} * \theta * (\gamma - 1) \end{aligned}$$

This shows that heterogeneous network energy has $\theta * (\gamma - 1)$ more virtual nodes than homogeneous networks with the energy $N * E_{ini}$.

HMCA algorithm is based on EEMC [2]. It also uses the idea of TRMRP of PAMC, described in [6]. PAMC uses the concept of Minimum Reachability Power (MRP). MRP can be considered as transmission range. Distance parameter $\sum 1/dis(n_v, n_{sink})$ of EEMC, is replaced by $\sum 1/P_v$, the Total Reciprocal of Minimum Reachability Power (TRMRP) for all active nodes. $TRMRP = \sum 1/MinPow_i$, where $MinPow_i$ is minimum power level required by the node i to reach its CH. PAMC protocol also assumes that for any power level L_i , there is corresponding transmission range R_i such that

$$R_i < R_j \quad \forall L_i < L_j$$

The operation of the algorithm can be divided into following two phases:

(i) Cluster set-up phase

Initially, for each round, all nodes are set to regular (non-CH). Cluster set-up phase is initiated by base station by sending a *Start* beacon. All nodes select its minimum power level needed to reach the base station. The process of MRP discovery to base station is done only once during network life time and cached for subsequent usage. As active nodes receive the *Start* message, they reply by sending their residual energy $E_u(t)$ and the MRP P_u to the sink node to indicate that level- l CH will now be selected. On receipt of these values, sink broadcasts a 'command' message with the values like total remaining energy of the network $\sum E_v(t)$ and TRMRP $\sum 1/P_v$. With help of these values each active node u sets its probability of becoming CH. Since network is heterogeneous, nodes having more energy have greater chances to become CH. The probabilities of becoming CH for normal nodes $p_1^n(u)$ and advance nodes $p_1^a(u)$ can be given by following formulas,

$$p_1^n(u) = N_{CH1}^{opt} \left[\varphi \frac{E_u(t)}{\sum_{v \in S(t)} E_v(t)} + (1 - \varphi) \left(\frac{1/P_u}{\sum_{v \in S(t)} 1/P_v} \right) \right], \text{ and}$$

$$p_1^a(u) = N_{CH1}^{opt} \left[\varphi \frac{E_u(t) * \gamma}{\sum_{v \in S(t)} E_v(t)} + (1 - \varphi) \left(\frac{1/P_u}{\sum_{v \in S(t)} 1/P_v} \right) \right].$$

Here, φ is a parameter deciding the weightage of energy factor and distance factor, and N_{CH1}^{opt} is the optimal number of CHs for level-1. Here $\sum_{v \in S(t)} E_v(t) = N * E_u(t) * (1 - \theta + \theta * \gamma)$ and $S(t)$ is assumed to be the active node set of the network at time t .

All level-1 CH send out a level-1 CH message associated with MRP. Any node that receives this message replies with a “join” message sending out its residual energy and MRP to the CH.

When level-1 CHs (denoted as n_{CH1}) receive this message, they can construct their active node-set, denoted as $S_i(t)$, and then broadcast a “command” message with following three values: total residual energy of the cluster, the TRMRP within the cluster of each node and the cardinality of its active set N_1 . After receiving this message each active node u will then set its probability $p_2^n(u)$ and $p_2^a(u)$ (for normal and advance node respectively) to become a level-2 cluster-head according to the following formula

$$p_2^n(u) = \sqrt{N_1} \left[\varphi \frac{E_u(t)}{\sum_{v \in S_1(t)} E_v(t)} + (1 - \varphi) \left(\frac{1/P_u}{\sum_{v \in S_1(t)} 1/P_v} \right) \right]$$

Here we make an additional assumption that after each round the percentage of normal nodes and super nodes selected for CH role remain constant i.e. value of θ remains unchanged.

$$p_2^a(u) = \sqrt{N_1} \left[\varphi \frac{E_u(t) * \gamma}{\sum_{v \in S_1(t)} E_v(t)} + (1 - \varphi) \left(\frac{1/P_u}{\sum_{v \in S_1(t)} 1/P_v} \right) \right]$$

In general, process of clustering may extend up to level- j . For any level- $(j-1)$, the elected CH will broadcast a level- $(j-1)$ CH message. Any node that receives this message replies with a “join” message sending out its residual energy and MRP to the CH.

When level- $(j-1)$ CHs (denoted as n_{CHj-1}) receive this message, they can construct their active node-set, denoted as $S_{j-1}(t)$, and then will broadcast a “command” message with following three values: total residual energy of the cluster, the TRMRP within the cluster of each node and the cardinality of $S_{j-1}(t)$ active set N_{j-1} . After receiving this message each active node u will then set its probability $p_j^n(u)$ and $p_j^a(u)$ (for normal and advance node respectively) to become a level- j cluster-head according to the following formula,

$$p_j^n(u) = \sqrt{N_{j-1}} \left[\varphi \frac{E_u(t)}{\sum_{v \in S_{j-1}(t)} E_v(t)} + (1 - \varphi) \left(\frac{1/P_u}{\sum_{v \in S_{j-1}(t)} 1/P_v} \right) \right]$$

$$p_j^a(u) = \sqrt{N_{j-1}} \left[\varphi \frac{E_u(t) * \gamma}{\sum_{v \in S_{j-1}(t)} E_v(t)} + (1 - \varphi) \left(\frac{1/P_u}{\sum_{v \in S_{j-1}(t)} 1/P_v} \right) \right]$$

The process of clustering will stop if a node has two or less number of nodes in its regular set.

(ii) Network operation phase

When level-T clustering topology is formed, the regular nodes start transmitting the sensed data to their CHs. Level-T CHs aggregate the sensed data and send it to level-(T-1) CHs and so forth. Finally, all level-1 CH transmit the aggregated data to sink node. Fig. 1 shows the wireless sensor network before first round (Fig. 1a) and final cluster formation (Fig 1b) after the first round in HMCA.

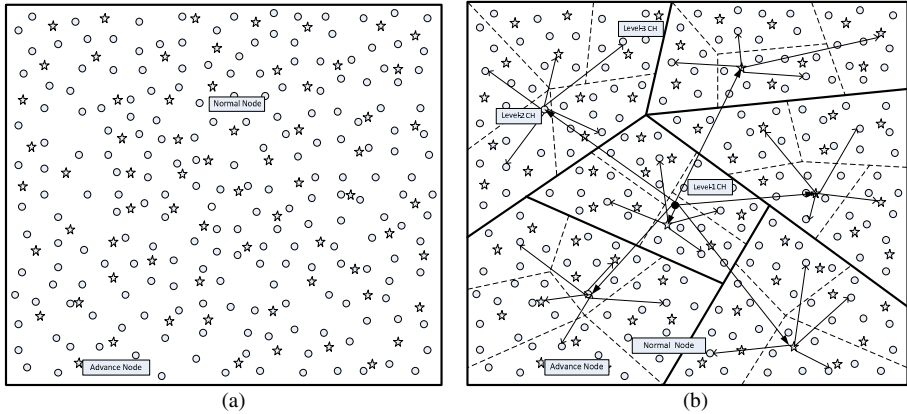


Fig. 1. Wireless sensor network before and after the first round

4 Simulation and Results

In this section we evaluate the performance of HMCA protocol. To validate the performance of HMCA, we have done simulation using ns-2 [8], a discrete event network simulator. Advance and normal nodes are randomly distributed over the field. We have compared the performance of HMCA with EEMC in terms of network lifetime. The basic parameters used are listed in Table 1.

Table 1. Simulation parameters

Parameter	Value
Network grid	100 × 100
Number of nodes	200-500
Base station position	50 × 50
ϵ_{fs}	10 pJ/bit/m ²
E_F	5 nJ/bit
E_{elec}	50 nJ/bit
Size of data packet	500 bits
Size of control packet	10 bits
Initial energy of normal nodes	1 J
Transmission range of sensor node	10 m

We used the idea of first node dead (FND) and half node dead (HND) to evaluate the network lifetime. FND and HND are number of rounds that elapse before first node and half of the nodes, respectively, run out of energy. For heterogeneous environment we have two types of nodes: normal nodes and advance nodes. We carried out the simulation for EEMC and HMCA for $\theta=0.2$ and $\gamma=3$. Fig. 2 and Fig. 3 show that HMCA outperforms EEMC in terms of network lifetime. This shows that death of first node and last node occurs earlier in EEMC.

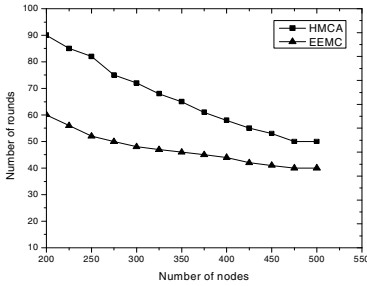


Fig. 2. The FND analysis

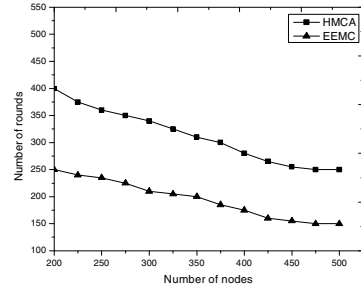


Fig. 3. The HND analysis

5 Conclusion

The wireless sensor networks have been designed to help in various monitoring applications. For real life deployments of large scale WSNs, multilevel clustering is the need for better connectivity. In this paper, we have proposed HMCA, a multilevel clustering scheme for HWSNs. Our proposed protocol is based on EEMC and uses the idea of minimum reachability power (MRP) for CH selection in heterogeneous environment. Simulation results show that, in presence of heterogeneity HMCA outperforms EEMC.

References

1. Corchado, J.M., Bajo, J., Tapia, D.I., Abraham, A.: Using heterogeneous wireless sensor networks in a Telemonitoring system for healthcare. *IEEE Transactions on Information Technology in Biomedicine* 14(2), 234–240 (2010)
2. Jin, Y., Wang, L., Kim, Y., Yang, X.: EEMC: An Energy-Efficient Multi-Level Clustering Algorithm for Large-Scale Wireless Sensor Networks. *Computer Networks* 52(3), 542–562 (2008)
3. Smaragdakis, G., Matta, I., Bestavros, A.: SEP: A stable election protocol for clustered heterogeneous wireless sensor networks. In: *Proc. of the International Workshop on SANPA 2004*, pp. 251–261 (2004)
4. Kumar, D., Aseri, T.C., Patel, R.B.: EEHC: Energy efficient heterogeneous clustered scheme for wireless sensor networks. *Computer Communications* 32(4), 662–667 (2009)
5. Qing, L., Zhu, Q., Wang, M.: Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks. *Computer Communications* 29(12), 2230–2237 (2006)

6. Soni, S., Chand, N.: Energy Efficient Multilevel Clustering to Prolong the Lifetime of Wireless Sensor Networks. *Journal of Computing* 2(5), 158–165 (2010)
7. Yarvis, M., Kushalnagar, N., Singh, H.: Exploiting heterogeneity in sensor networks. In: *IEEE INFOCOM* (2005)
8. VINT Project. The ucb/lbnl/vint network simulator-ns,
<http://www.isi.edu/nsnam/ns>

Anomaly Detection in Ethernet Networks Using Self Organizing Maps

Saroj Kumar Panigrahy, Jyoti Ranjan Mahapatra,
Jignyanshu Mohanty, and Sanjay Kumar Jena

Department of Computer Science and Engineering
National Institute of Technology Rourkela, 769 008, Odisha, India
{panigrahys, skjena}@nitrkl.ac.in

Abstract. Anomaly detection attempts to recognize abnormal behavior to detect intrusions. We have concentrated to design a prototype UNIX Anomaly Detection System. Neural Networks are tolerant of imprecise data and uncertain information. A tool has been devised for detecting such intrusions into the network. The tool uses the machine learning approaches and clustering techniques like Self Organizing Map and compares it with the K -means approach. Our system is described for applying hierarchical unsupervised neural network to intrusion detection system.

Keywords: Intrusion detection, anomaly detection, self organizing map, neural networks.

1 Introduction

A secure computer network is one that assures data confidentiality, data and communications integrity and protection from denial of service (DOS) attacks [1]. The paradigm for securing computer networks was eventually replaced by the notion of intrusion detection [2]. There are many types of intrusion detection systems, but most can be classified in one of two ways [1]. First, an intrusion detection system can be classified based on the data source that it uses. A host-based intrusion detection system uses the audit trails of the operating system as a primary data source. For example, it may use records of user sessions to detect particular sessions that constitute an intrusion. A network-based intrusion detection system, on the other hand, uses network traffic information as its main data source. An example would be a system that uses TCP header information.

When analyzing any intrusion detection system, three factors must be considered—efficiency of the system, timeliness of detection and accuracy of detection [2]. The basic idea behind the system is that hierarchies of self organizing maps (SOM) take on a divide and conquer approach to concisely model the normal behavior of the system. Given the model of normal behavior, running data that corresponds to some suspicious behavior through the system will then exhibit some telltale signs that can be used to raise an alarm. The system described herein is a network based anomaly detection system that uses TCP dump dataset [3].

2 Advantages of SOM over K -means Approach

Authors are ambiguous over which is the best method for implementing the anomaly detection technique. Some support the SOM approach and some favour the K -means approach. But one major area of concern is not the difficulty of run time complexity but the difficulty of implementing it over dataset. The K -means approach needs a mean centroid to be defined at the outset of the run. But network traffic data is variable and deviates randomly over time. So, any randomly generated initial mean value is difficult to implement on variable network traffic. On the other hand SOM learns itself according to given data. Also k -means approach suffers from the problem of local optima. As the initial centroid value and number of clusters is chosen, it might be away from the optimal centroids and the end result for SOM is better than K -means. Search space is better explored by SOM due to the effect of the neighbourhood parameter which forces units to move according to each other in the early stages of the process. K -means gradient orientation forces a premature convergence which, depending on the initialization, may frequently yield local optimum solutions.

3 Implementation Details

This section describes about the dataset, SOM architecture, data preprocessing, SOM training, calibrating The code book vectors, and Running the dataset referring with code book vector.

Dataset

The dataset available for constructing the system consisted of nearly five million connections of labeled training data and two million connections of test data. The connections were in chronological order. Each connection was described by 41 features. The features can be categorized as— *basic TCP features, content features, time based traffic features, and host based traffic features* [4]. A connection in the training data was either a normal connection or was one of 24 different attack types. Each connection was either normal or fell into one of the categories of attacks— *remote-to-local, user-to-root, denial-of-service, probing*.

From the available features, six were selected for use in the system— *duration, protocol type, service, flag, destination bytes, and source bytes*. Three of these features— *duration, destination bytes, and source bytes* had continuous values and *Protocol type, service, and flag* had discrete values. It should be noted that the entire dataset consisting of the seven million connections was not used in constructing the system. Only a 10% dataset from among the connection was used in order to make the training computationally feasible and most traffic has a typical pattern. Capturing the pattern of the traffic once is sufficient than doing it repeatedly. The 10% dataset represented the whole traffic connection for the training purpose. The dataset was extracted from the KDD Cup dataset which consisted of TCP dump data of DARPA Intrusion Detection Evaluation [3].

SOM Architecture

A SOM architecture with a single level was used [2]. Such an architecture was shown to be effective for the purpose of intrusion detection. The algorithm was fed with the above six parameters chosen. The data was preprocessed to get into a form that was program readable. To extract the TCP dump data, a network sniffer was used. The sniffer was placed on a central hub through which all traffic is routed so that it can capture all packets in promiscuous mode. It is a static dataset and used to standardize the algorithm for use on a more dynamic traffic data. The single level map model the behavior of the computer network with respect to time and given feature. The data flow in SOM is shown in following Figure.



Data Preprocessing

The first step in preprocessing the data involved removing all the attack connections from the training dataset, leaving only the normal ones. Care was taken to preserve chronological order. The second step in preprocessing the data involved extracting each feature from this file. This resulted in a sequence of feature values, one per feature. Next, because three of the six features consisted of discrete string values, a format that cannot be fed directly into the SOM, these features had to be enumerated. The result of the extraction and enumeration was six sequences of numbers, with each sequence corresponding to a feature. The n th entry in all of these sequences corresponded to the six features for the n th connection in the dataset. As is, if the values in each sequence were fed to the maps, no temporal relationship would have been encoded. In order to encode ordering and frequency relationships in the patterns that the maps would see, a FIFO buffer was used [5]. For a buffer of size n , the basic form of this algorithm is of following form:

1. The values of the sequence are fed into the buffer in chronological order.
2. Once the n positions of the buffer are filled, a pattern is generated.
3. When the next value in the sequence is observed, the oldest value currently in the buffer is discarded, the remaining values in the buffer shift by one so that the vacated position is filled and the next value is placed in the empty location. This generates the next pattern.

Training the SOM

The map was trained on a block of 15,000 consecutive connections, a fraction of the total dataset available after the first preprocessing stage. Although training on more patterns would allow the system to model a wider range of normal behavior, it would make the training of the maps difficult given the available time line. The maps were trained using C. The result was a 10×10 map. Training uses all the mathematical calculations as described in [6]. For an input pattern

given to the map, its distance to each mapping unit is found out and normalized. In this way, patterns close to map units yield a normalized distance close to one, and patterns far away from it yield a normalized distance close to zero. This normalization was done so that the values for all the features would have the same range. Otherwise, certain features would dominate the distance measure used in training, and thus the training of the map, simply because they had a larger range and not necessarily because they were more significant. For each pattern, the normalized distance to each center in the map was recorded, resulting in a six dimensional vector for each map. The vectors for all the maps were then concatenated to form one vector of dimension 100.

Calibrating the Code Book Vectors

The code book vectors were calibrated with carefully selected input patterns so that attack patterns are not there [6]. These normal input data patterns map to some of the mapping units and these mapping units are labeled as normal. So the code book entries represent the anomalous as well as the normal patterns and are labeled. During testing any pattern that doesn't match to these units are termed anomalous and an alarm is raised.

Running the Dataset

The dump file from KDD dataset is run referring the code book vectors and the output is generated along with labels signifying which input patterns were termed as anomalous. The false positive and false negative rates are calculated based on the output type, whether it is anomalous or not, and the input pattern, whether it was actually an attack packet.

4 Results

The algorithm was run on the test dataset and according to the given architecture. First, the test data was preprocessed, the SOM was trained with training data, the code book entries were labeled by calibrating with normal connections, and then the resulting code book entries were used for running the algorithm. Using the SOM implementation, the following results were obtained.

- Dimension of grid used: 10×10
- Samples Taken for training: 15,000
- Samples taken for testing: 13,500
- Attacks detected: teardrop, portsweep, ipsweep, backdoor, nmap, neptune, satan, phf, warezmaster
- Attacks not detected: pod, buffer_overflow, guess_passwd, imap, ftp_write, toolkit
- Dataset: KDD 10% unlabelled training dataset and 10% labeled testing dataset
- False Positive rate: $2/13500 = 1.07\%$
- False Negative rate: $145/13500 = 0.015\%$

The efficiency of the anomaly detection tool is reflected from the false positive rate, false negative rate and the number of attacks that were detected. In this tool, false positive rate was found to be very low. On the other hand, in the small dataset taken, false negative rate was also found to be low. But the system is inefficient because a number of attacks were not detected— pod, buffer_overflow, guess_passwd, imap, ftp_write, toolkit. The low level of false negative rate was due to the reason that these are not DoS type of attacks and the six parameters we chose for implementing the algorithm were identical in all respects for these attack packets and the normal packets. As these packets are encountered less in number in the traffic, the numerator value in calculation becomes small, and hence the low false negative rate.

A major analysis would be around the detection of the mapping units of the 10×10 grid. The points of the grid where most normal packets match, i.e., the frequency with which the packets map to particular grid units. Figure 1 shows the number times each node in the top level map was the BMU for the normal training data. Clearly, some nodes are BMUs more frequently than others, but most of the nodes receive their fair share of hits. However, nodes 10, 17, 18, 26, 27, 28, 29, 38, 48, 55, 57, 58, 59, 60, 66, 67, 68, 69, 70, 79, 80, 89, 90, 99, 100 stand out because they receive relatively few hits. Thus, these nodes could be considered to be associated with abnormal behavior.

The performance of this system is comparable to that of the systems participating in the DARPA Intrusion Detection Evaluation 1999. The best system in

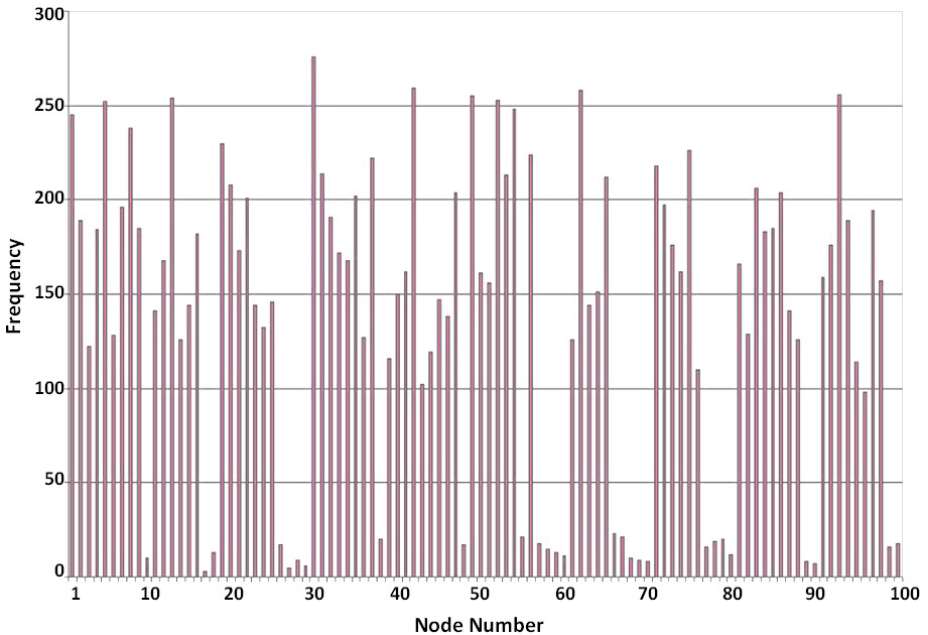


Fig. 1. Hits per node of normal training data

the evaluation had an overall false negative rate of about 0.33 and an overall false positive rate of 0.0002. This system used all the available TCP connection features, and was trained on the entire available training data set which is shown in Table 1. Given that the system presented in this paper used only a fraction of this information, its performance is solid.

Table 1. Fraction of Data used

	Number of Connections Used	Fraction of Total
Training SOM	15,000	0.05
Labeling SOM	4,94,021	0.1
Testing SOM	3,11,029	0.1

5 Conclusions

A network based anomaly detection system that uses a hierarchy of SOMs was presented. The system was found to detect just over 60% of the attacks with a manageable rate of false alarms. Although the results of this work should be interpreted with caution, it is suggested that the system presented performs comparably to some of the better systems that took part in the DARPA Intrusion Detection Evaluation. The system was not tested on the full test dataset, i.e., it may not have encountered some of the more difficult attacks but it was also never trained on the full training dataset, meaning that it may not have had a chance to learn the full range of normal behavior.

References

1. Mukherjee, B., Heberlein, L.T., Levitt, K.N.: Network intrusion detection. *IEEE Network* 8(5), 26–41 (1994)
2. Lichodziejewski, P., Zincir-Heywood, A.N., Heywood, M.: Dynamic intrusion detection using self-organizing maps. In: *Proceedings of the 14th Annual Canadian Information Technology Security Symposium*, Ottawa, Canada (May 2002)
3. KDDCup: The third international knowledge discovery and data mining tools competition (May 2002), <http://kdd.ics.uci.edu/databases/kdd99cup/kdd99cup.html>
4. Lee, S.C., Heinbuch, D.V.: Training a neural network based intrusion detector to recognize novel attacks. *IEEE Transactions on Systems, Man and Cybernetics* (July 2001)
5. Lee, W., Stolfo, S.J., Chan, P.K., Eskin, E., Fan, W., Willer, M., Hershkp, S., Zhang, J.: Real time data mining based intrusion detection. In: *Proceedings of DISCEX II* (June 2001)
6. Kohonen, T.: *Self Organizing Maps*, 3rd edn. Springer, Heidelberg (2001)

A Heuristic Multi Criteria Routing Protocol in Wireless Sensor Networks

Alireza Shams Shafigh¹ and Marjan Niyati²

¹ Computer Engineering Department, Damavand Branch, Islamic Azad University

² ICT Department, Iran-Transfo Factory
Tehran, Iran

Ali_Shams@damavandiau.ac.ir
m.niyati@iran-transfo.com

Abstract. Most existing routing techniques are designed to reduce routing cost by optimizing one goal. In wireless sensor networks one parameter cannot satisfy all requirements of these networks. Therefore, it's required to consider multiple parameters simultaneously in the routing process for improving performance of the routing protocols. However, the multiple-criteria routing process leads to some increases in complexity of the routing protocol. Recently heuristic methods specially the learning automata were used in the optimization problem. Hence, in this paper, we propose a multi-criteria routing method which uses learning automata to choose the optimum route between the available routes. Results of our simulation present high improvement in compared to the other routing methods.

Keywords: wireless sensor networks, learning automata, quality of service, multi-criteria routing.

1 Introduction

Wireless sensor network (WSN) [1], [4] has attracted upon intensive research recently, due to its large range of applications. In the most scenarios, sensor networks are modeled as data acquisition systems which are a collection of sensor nodes gathering environmental information. This information is aggregated and routed via other sensors to a powerful base station to be processed. The data packets are relayed from the data sources to the base station. Unlike network with powerful nodes and stable links, ad hoc wireless sensor networks feature great dynamics like as unreliable communication, and frequent topology change. Hence it is reasonable to adopt reactive routing protocol that creates a routing path on demand. Reactive routing protocol enables the following scenario: starting from any node, one selects a successor from the neighbors of current node. By repeatedly jumping from current node to the successor, the base station is finally reached. Due to the fact that the sensor nodes are battery-powered devices, prolonging the autonomous lifetime of the network is a challenging problem. In designing energy-efficient routing protocols, the various features of sensor networks lead to a set of optimization problems: Routing path length, Link reliability, minimum bandwidth, lowest end to end delay and low losses. Most existing routing techniques were designed to optimize one of these goals.

In real scenarios however, these factors are usually in conflict, and influence the routing performance in a complex way, leading to the need of a more sophisticated routing scheme that makes correct trade-offs.

The various features of wireless sensor networks lead to a set of optimization problems in designing energy-efficient routing protocols. Most existing routing techniques are designed to optimize one of these goals. For example, GPSR [9] and [7] are proximity based routing strategies that aim at finding the shortest routing path; [14] proposed a routing scheme that takes into account the residual energy of sensor nodes, in order to balance the load; Directed Diffusion [8] is a fault tolerant routing strategy that is robust against unstable communication. To the best of our knowledge, no existing routing scheme takes in consideration all these optimization goals together.

The idea of applying reinforcement learning to routing in networks was firstly introduced in [3], [15], [16] in static packet switched network. They have shown that the Q-learning based routing approach is able to compete with the shortest path algorithms, without prior knowledge regarding the network topology. Following their prior work, Q-learning is applied to routing in ad-hoc networks [6] and optimizing the tradeoff between routing and compression [12].

In this paper, we present a novel routing scheme called *QLABER*¹ which enables the sensor nodes to efficiently learn an *optimal* routing strategy, depending on multi-goals such as end to end delay, transmission rate or bandwidth, power level, hop count and loss rate. The remainder of the paper will be organized as follows: in the section 2, we describe in detail our routing scheme based on reinforcement learning; the experimental results are shown in the Section3; the paper is ended with a conclusion and future work in the Section4.

2 The Proposed Approach

In this paper, we propose a heuristic method which considers five parameters in its routing process. These parameters include bandwidth, energy level of node, number of hops from the sink, node's response as delay and loss rate. It's almost impossible to optimize the route selection process by mathematic models. Thus, we develop a new heuristic method that there are not any difficulties in its implementation. Our main idea originates from LABER protocol that includes two parameters in the routing process. The proposed method has two phases: the first phase is the creating route tables and the second phase is the routing and the learning process. In the following sections, we describe these two stages in the details.

2.1 Creating Routing Tables

At the first, the sink node floods a packet to be created route tables in all sensor nodes. The flood packet includes all information which has been shown in the figure1. This information is modified by the sender of this packet. Every node which receives this packet adds a new entry to its route table. After recoding information, node replaces its information to packet and rebroadcasts it to its neighbors.

¹ Quality of service Aware LABER.

Sender address
Power level of sender
Node delay
Hop count
Transmission rate

Fig. 1. Structure of the FLOOD packet

Every entry in node’s route table has the following information that has been shown in the figure2. This information is initiated in the first phase and updated in the routing and the learning phase.

address
Power level
delay
Hop count
All sent packets
All loss packets
Transmission rate
Last update time
Probability of selection

Fig. 2. Entry of route table

When a node receives a FLOOD packet, it runs the following function to update its route table. Each entry in a route table has a probability that presents its selection’s probability to send data packets. If a sensor node receives only one FLOOD packet, there is one entry in its route table. Therefore, the selection’s probability of this route is one; on the other hand, the selection’s probabilities of routes in a route table computed by the formula (1).

```

For each node (::)
{
  If Receive a new (FLOOD packet)
  {
    Extract information
    Add new entry to route table
    Update probabilities of routes in route table by
      formula (1)
    If time to live of (Flood packet) was not expired
    {
      Replace this node information to (FLOOD
        packet)
      Send (FLOOD packet) to neighbors
    }
  }
}

```

In the formula (1), m is number of available routes in a route table. Also P is the selection’s probability of route_i according to hops count (num Hop_i), energy level

(energy_i), node's response delay (delay_i), loss rate (loss_i) and node's transmission rate or bandwidth (bandwidth_i).

$$\forall i \leq m \quad P_i = h_1 \frac{1}{\sum_{i=1}^m \frac{1}{numHop_i}} + h_2 \frac{energy_i}{\sum_{i=1}^m energy_i} + h_3 \frac{1}{\sum_{i=1}^m \frac{1}{delay_i}} + h_4 \frac{bandwidth_i}{\sum_{i=1}^m bandwidth_i} + h_5 \frac{1}{\sum_{i=1}^m \frac{1}{loss_i}} \quad (1)$$

Node's response delay is determined by delay from the sending time of a data packet to the receiving time of acknowledge packet. This parameter presents how much a neighbor node is busy. Loss rate is another parameter that is considered in the route selection process. In this paper, we use transmission rate as bandwidth. We can determine the impact of every parameter in the formula (1). Values of h_j must be selected such as h₁+h₂+h₃+h₄+h₅=1. As a result of the formula (1), one probability is computed for every entry in the route table to show how much is possible to be selected this route. Efficient parameters causes to a great increase in probability of a route.

2.2 Learning and Routing

When a sensor node is going to send a data packet, it selects one of the best routes in its route table in regard to values of route's probabilities. When It forwards data packet to the sink node via the selected node (highest probability), it set a timer and waits for the acknowledge packet. The figure3 presents structure of the Ack packet. This packet carries only information about power and bandwidth; other parameters are calculated by expiration of timer and receiving a correct ACK packet. There are two conditions after sending a data packet. In the first condition, we receive an ACK packet that presents the data packet was correctly delivered to next node. In the second situation, timer is expired without any acknowledge packet. Therefore, it's supposed that one data packet was lost. Thus, node's loss rate presents how much of data packets were acknowledged by that node. The loss rate is computed by the formula (2).

$$loss \ rate \ = \ 1 \ - \ \frac{(number \ of \ ACKs)}{number \ of \ DATAs} \quad (2)$$

To compute response time for a node, the formula (3) is used. T_e presents expiration time of timer; T_a define receiving time of acknowledge packet; T_s is sending time of data packet.

$$response \ time \ = \ T_e \ or \ T_a - T_s \quad (3)$$

All new facts include inside of the ACK packet and those are computed by formulas (2) and (3) are used to decide about rewarding or penalizing.

Sender address
Power level of sender
Transmission rate

Fig. 3. Structure of the ACK packet

When nodes receive an ACK packet, it extracts packet's information to update route table. An Ack packet specify how much efficient has been the current node's action. In fact, ACK packet is a reaction from environment of LA or especially from next node. Another reaction is committed when node's timer is expired. To do an action, nodes select one path that has the highest probability from their route table. Thus, it's time of judgment when a node receive an ACK or was expired its timer. LA fines all paths that have low efficiency to use for sending data packets. Meanwhile, it reward those paths has more resources. LA's main goal is to heuristically choose the best paths to forward data packets via them. The following pseudo code presents action and reaction of one typical node.

```

For each node (i)
{
  If Generate (DATA packet)
  {
    Select one Path// with the aid of Learning
      Automata
    Transmit (DATA packet) on the selected path
    Wait for ACK
  }
  If received (DATA packet)
  Transmit (ACK packet) to sender node
  If not Destination
  {
    Select one Path // with the aid of Learning Automata
    Transmit (DATA packet) on path Wait for ACK
  }
  If received (ACK packet)
  If ACK== "reward"      then      // a good path
    Reward the path and update table
  If ACK== "Penalty"    then      //a bad path
    Penalize the path and update table
  If don't received after of waiting for ACK then
    Suppose a packet was loss
  }
}

```

As it was mentioned, we consider five parameters to specify how much a path is efficient. The formula (4) determines impact's value ($\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$) of different issues in selecting process.

$$Q = +\alpha_1 * enrg + \alpha_2 * bw + \alpha_3 / del + \alpha_4 / loss + \alpha_5 / hop \quad (4)$$

In the formula (4), enrg presents power level of node, bw is transmission rate of a node, del is response time, loss shows value of loss rate and hop is number of hops from sink. After computing value of Q, node's LA decides about its previous action. In the formula (4), $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 = 1$. The following pseudo code describes how a LA updates its probabilities in route table.

The following algorithm is run when an ACK packet is received or a timer is expired. In the pseudo code, $Q_{average}$ is average of all Q for the existence routes in

route table. We define two thresholds th_1 and th_2 to evaluate the reactions. These parameters are set to 0.4 and 0.75 in our simulation. To fine an action, its selection's probability is reduced by β and other existence routes in route table are increased by β/n (n is number of routes in route table). In the formula (5), the value of β is computed.

```

For each node ( :: )
{
  If Receive (ACK packet) or expire its receiving time
  {
    If Receive (ACK packet)
    {
      Extract information on ACK packet
      Compute new value of Q
    }
    Else
    {
      Compute new value of Q with the updated loss
      rate and other old information in route table
    }
    Compute average quality for all routes in route
    table as  $Q_{average}$ 
    If ( $Q < th_1$ )
      Punish the selected route with value of  $\beta$ 
    Else if ( $(Q > th_1) \ \&\& \ (Q < th_2)$ )
      Reward the selected route with value of  $\alpha/2$ 
    Else
      Reward the selected route with value of  $\alpha$ 
  }
}

```

In our method, each LA learns to use a better route by continuously rewarding or penalizing. The formula (5) presents value of punishment for a wrong selection. To determine value of β , average of energies (avg_{enrg}), maximum of bandwidths (max_{Bw}), maximum of hop counts (max_{Hop}), maximum of delays (max_{Delay}) and maximum loss rates (max_{Loss}) in route table must be computed. In the formulas (5) and (6), $enrg_i$ is power level, bw_i is bandwidth, Hop_i is hop count, $delay_i$ is response delay and $loss_i$ is loss rate for a typical node,

$$\beta = \frac{((avg_{enrg} - enrg) + (max_{Bw} - bw_i)) + Hop_i + delay_i + loss_i}{avg_{enrg} + max_{Bw} + max_{Hop} + max_{Delay} + max_{Loss}} \quad (5)$$

To reward an action, the value of α is added to its selection's probability in route table (the formula (6)). For other routes value of α/n is decreased from their probabilities.

$$\alpha = \frac{(enrg_i + bw_i) + (max_{Hop} - Hop_i) + (max_{Delay} - delay_i) + (max_{Loss} - loss_i)}{avg_{enrg} + max_{Bw} + max_{hop} + max_{delay} + max_{Loss}} \quad (6)$$

3 Experimental Results

In order to evaluate the performance of our approach, we simulate several scenarios using glomosim simulator in according to the following parameters (the table 1):

Table 1. Enviroment of simulation

Simulation time	600 second
Simulation area	1000m*1000m
mobility	Random waypoint with 0 pause time

To present our method's performance, we assign to each node a different physical property; For example, each node has a random bit rate between 2Mbps to 10 Mbps. This bite rate is used as bandwidth in our proposed method. Another critical resource in MANETs is power level (battery). Any node has a battery which used to send and receive packets. The default values for formulas are determined by the simulation results. We evaluate our method which we call it as QLABER with LABER protocols to present achievements of QLABER. To evaluate the results of the proposed method, the following parameters have been studied: *End to end delay*, *Packet delivery ratio*, *Consumption power*. To present the performance of our method in related to LABER, in all scenarios of this paper, we use sensor nodes which have very different physical properties.

3.1 Impact of Number Nodes in Network

In the first scenario, we explore the impact of number nodes to the proposed method's performance in a wireless sensor network. Our method could adapt self with different situation. Indeed, it tries to use the better routes in forwarding data packets. To make an optimum decision, LA has to do many try and errors by sending data in different routes. The figure (4) presents the average delay of our method in compared to LABER. Our method has lower end to end delay then LABER in a high-dense network. Although a high-dense network results in some problems like as high collision, it introduces more resources to its members. Thus, our method efficiently uses this opportunity to bring more improvement to its services.

LABER heuristically selects the better routes in regard to energy and hop count. These factors are not only included in our method includes but also the delay, loss rate and bandwidth. So our method could achieve the better results. The figure 5 shows that QLABER remarkably has high packet delivery ratio in regard to LABER. Since our method delivers more data packets, it efficiently uses the network's resources especially power of nodes. Besides, our method sends data packets via the more reliable than LABER does. Thus, while LABER delivers very low number data packets, it uses a great deal of powers. It's obvious that our method's improvement originates from multi-criteria routing.

As mentioned in above, a high-dense network proposes more resources. Therefore, our method uses this facility to achieve the better results. In contrast, LABER couldn't efficiently use it.

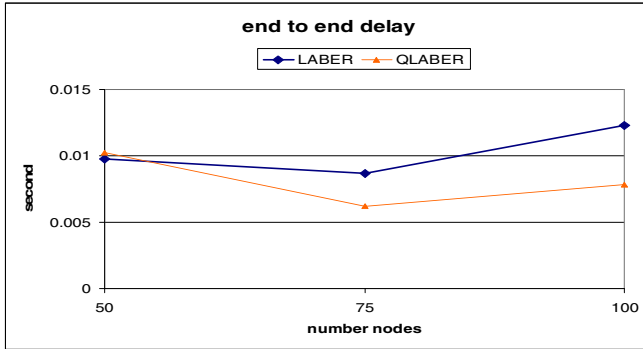


Fig. 4. End to end delay (second)

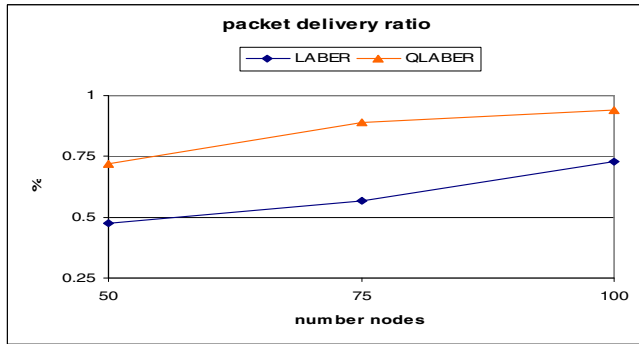


Fig. 5. Packet delivery ratio

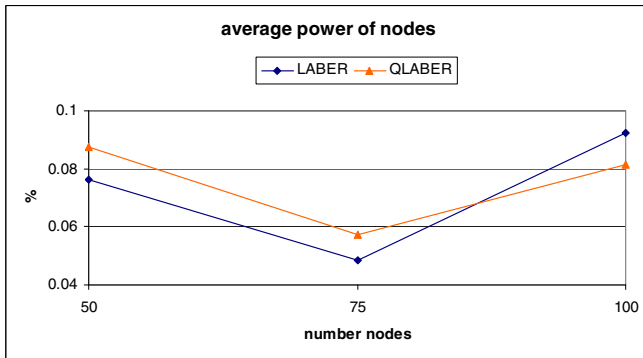


Fig. 6. consumed power (miliwatts)

3.2 Impact of Input Traffic

Learning automata tries many different routes to select the best route based on the quality parameters. Each LA updates its route table and probabilities with sending a data packet and receiving acknowledge packet. Therefore, when a node sends more data packets, the frequency of route table's update is increased. The figures 7-9 present all results for the second scenario. In this scenario, we change the interval time between two continuous data packet sending in the originator. Obviously, our method due to including more variables in its decision achieves very remarkable improvements in compared to LABER.

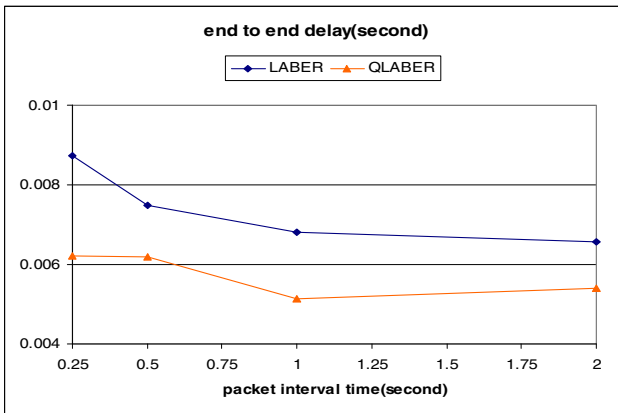


Fig. 7. End to end delay (second)

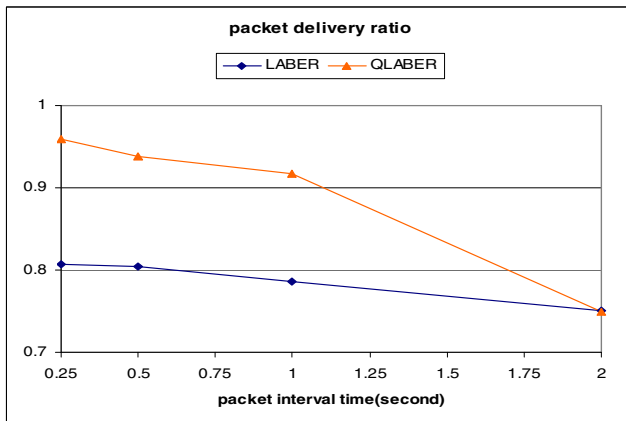


Fig. 8. Packet delivery ration (%)

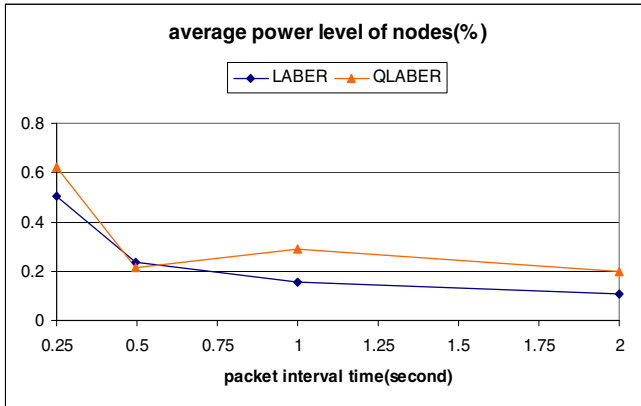


Fig. 9. Consumed power (milliwatt)

When number of action and its related reaction increases, LA gains more accurate knowledge about their environment. In all figures, increase of packet interval results in decrease the input traffic.

3.3 Impact of Loss Rate

In this scenario, our method tries to avoid from nodes which have high loss rate. There are some nodes in network with high loss rate; this means these nodes randomly neglect to forward data and ACK packets. In this scenario, number of malicious nodes is determined by a percent. The figures 10-12 show the result of this scenario when percent of the malicious nodes changes from 0.1 to 0.5 of all nodes in network. .LABER has no plan for this threat in real network; so fails to overcome to this problem. Indeed, in a harsh environment, our method can be used without any severity.

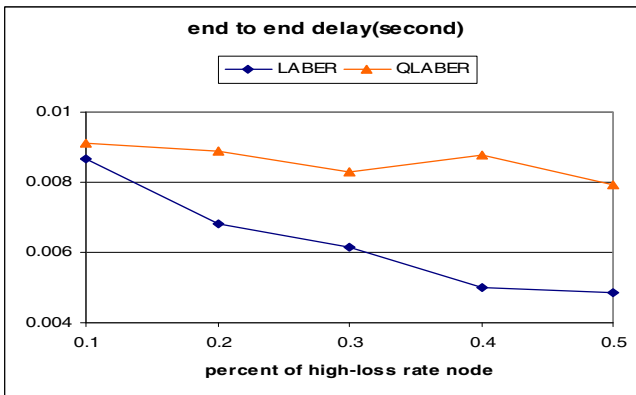


Fig. 10. End to end delay (second)

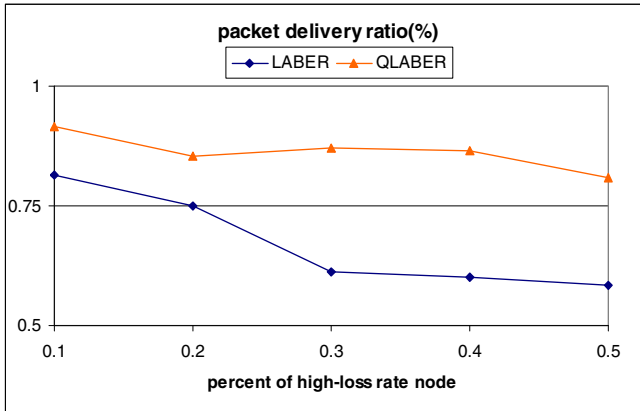


Fig. 11. Packet delivery ration (%)

In the figure 10, our method due to delivering a great deal of data packets has higher end to end delay than LABER which blindly send packet to high-loss rate nodes. Our method calculate number of lost acknowledge and data packets by every neighbor node; so it efficiently use this information to select a better route. The figures 11 and 12 present packet delivery ratio and average power level of nodes in the end of simulation.

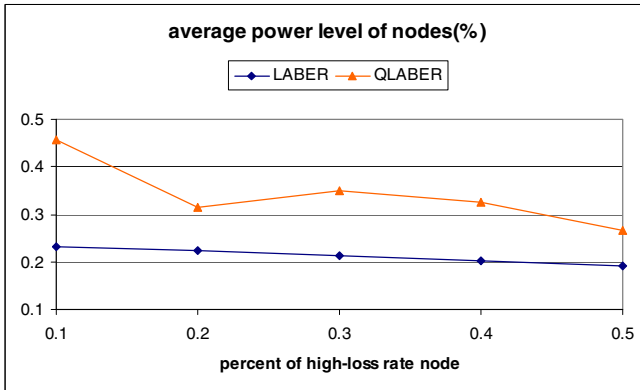


Fig. 12. Consumed power (milliwatt)

Our method completely has more improvement in regard to many criteria due to its more accurate making decision.

4 Conclusion and Future Works

We present a novel routing scheme which enables the sensor nodes to efficiently learn an *optimal* routing strategy. The proposed method has no complexity to implement, while it efficiently improves the performance of multi-criteria routing methods in WSN. Our simulation’s results show that QLABER is very efficient routing protocol

in heterogeneous networks. Our future work is to compare QLABER with other heuristic methods as example fuzzy logic, ant colony, etc.

References

1. Akyildiz, F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. *Computer Networks* (38), 393–422 (2002)
2. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks. *IEEE Communication Magazine* 40, 102–114 (2002)
3. Janakiram, D., Venkateswarlu, R., Nitin, S.: A survey on programming languages, middleware and applications in wireless sensor networks. IITM-CSE-DOS-2005-04 (2005)
4. Estrin, D., et al.: Embedded Everywhere: A research agenda for network systems of embedded computers. Computer Science and Telecommunication Board (CSTB) Report. National Academy Press (2001)
5. He, T., Stankovic, J.A., Lu, C., Abdelzaher, T.F.: Speed: A stateless protocol for realtime communication in sensor networks. In: IEEE International Conference on Distributed Computing System, ICDCS 2003, pp. 46–55 (2003)
6. Felemban, E., Lee, C.-G., Ekici, E.: Mmspeed: Multipath multispeed protocol for qos guarantee of reliability and timeliness in wireless sensor networks. *IEEE Transaction on Mobile Computing* 5(6), 738–754 (2006)
7. Zeng, K., Ren, K., Lou, W., Moran, P.J.: Energy aware efficient geographic routing in lossy wireless sensor networks with environmental energy supply. *Wireless Networks* 15(1), 39–51 (2009)
8. Chipara, O., He, Z., Xing, G., Chen, Q., Wang, X., Lu, C., Stankovic, J., Abdelzaher, T.: Real-time ower-aware routing in sensor networks. In: Proceeding of the IEEE International Workshop on Quality of ervice, IWQoS (2006)
9. Lim, T.L., Gurusamy, M.: Energy aware geographical routing and topology control to improve network lifetime in wireless sensor networks. In: IEEE International Conference on Broadband Networks (BROADNETS 2005), pp. 829–831 (2005)
10. Esnaashari, M., Meybodi, M.R., Sabaei, M.: A novel method for QoS support in sensor networks. In: CSICC 2007, pp. 740–747 (2007)
11. Farajzadeh, N., Meybodi, M.R.: Learning automata-based clustering algorithm for sensor networks. In: Proceedings of CSICC 2007, pp. 780–787 (2007)
12. Gholipour, M., Meybodi, M.R.: LA-Mobicast: A learning automata based distributed protocol for mobicast in sensornetworks. In: Proceedings of CSICC 2007, pp. 1154–1161 (2007)
13. Narendra, K.S., Thathachar, M.A.L.: Learning Automata: AnIntroduction. Prentice Hall, Englewood Cliffs (1989)
14. Beigy, H., Meybodi, M.R.: A mathematical framework for cellular learning automata. *Advances on Complex Systems* 7(3-4), 295–320 (2004)
15. Thathachar, M.A.L., Sastry, P.S.: Varieties of learning automata: An overview. *IEEE Transaction on Systems, Man, and Cybernetics-Part B: Cybernetics* 32(6), 711–722 (2002)
16. Intanagonwivat, C., Govindan, R., Estrin, D., Heidemann, J., Silva, F.: Directed diffusion for wireless sensor networking. *Proceedings of the IEEE/ACM Transactions on Networking (TON)*, 2–16 (2003)
17. Janakiram, D., Venkateswarlu, R., Nitin, S.: A Survey on Programming Languages, Middleware and Applications in Wireless Sensor Networks. In: Proceedings of the IITMCSE- DOS-2005-04 (2005)

Framework and Implimentation of an Agent Based Congestion Control Technique for Mobile Ad-hoc Network

Sarita Singh Bhadauria¹ and Vishnu Kumar Sharma²

¹ Department of Electronics
MITS Gwalior (M.P.) India
saritamits61@yahoo.co.in

² Department of CSE
JUET, Guna (M.P.)
vishnusharma97@gmail.com

Abstract. In Mobile Ad hoc Networks (MANETs) obstruction occurs due to the packet failure and it can be successfully reduced by involving congestion control scheme which includes routing algorithm and a flow control at the network layer. In this paper, we propose to framework and implimentation an agent based congestion control technique for MANETs. In our technique, the information about network obstruction is collected and disseminated by mobile agents (MA). The nodes are classified into four categories based on its traffic class. The MA measures the get in line length of the various traffic classes and the channel disagreement and estimates the total congestion metric to find the minimum congestion level in the network. The congestion metric is applied in the routing protocol to select the minimum congested route. By simulation results, we show that our proposed technique attains high delivery ratio and throughput with reduced delay when compared with the presented technique.

Keywords: Mobile Ad hoc Networks (MANETs), mobile agents (MA), TCP, EDCA Mechanism of 802.11e, CBR, FTP, MAC Layer.

1 Introduction

1.1 Mobile Ad Hoc Networks

The mobile ad hoc network is capable of forming a temporary network, without the need of a central administration or standard support devices available in a conventional network, thus forming an infrastructure-less network. In order to guarantee for the future, the mobile ad hoc networks establishes the networks everywhere. To avoid being an ideal candidate during rescue and emergency operations, these networks do not depend on the irrelevant hardware. These networks build, operate and maintain with the help of constituent wireless nodes. Since these nodes have only a limited transmission range, it depends on its neighboring nodes to forward packets [1].

1.2 Congestion Control in MANETs

Congestion takes place in MANETs with limited resources. In these networks, shared wireless channel and dynamic topology leads to interference and fading during packet transmission. Packet losses and bandwidth degradation are caused due to congestion, and thus, time and energy is wasted during its recovery. Congestion can be prevented using congestion-aware protocol through bypassing the affected links [2]. Severe throughput degradation and massive fairness problems are some of the identified congestion related problems. These problems are incurred from MAC, routing and transport layers [3].

Congestion control is the major problem in mobile ad hoc networks. Congestion control is related to controlling traffic entering into a telecommunication network. To avoid congestive collapse or link capabilities of the intermediate nodes and networks and to reduce the rate of sending packets congestion control is used extensively [4]. Congestion control and reliability mechanisms are combined by TCP to perform the congestion control without explicit feedback about the congestion state and without the intermediate nodes being directly interrupted [4]. Their principles include packet conservation, additive increase/multiplicative decrease in sending rate, stable network. End system flow control, network congestion control, network based congestion avoidance, and resource allocation includes the basic techniques for congestion control [5].

2 Related Work

Wei Sun et al [8] have compared the general AIMD-based congestion control mechanism (GAIMD) with Equation-based congestion control mechanism (TFRC TCP-Friendly Rate Control) over a wide range of MANET scenario, in terms of throughput fairness and smoothness. Their results have shown that TFRC and GAIMD are able to maintain throughput smoothness in MANET, but at the same time, they require only a less throughput than the competing TCP flows. Also their results show that TFRC changes its sending rate more smoothly than GAIMD does, but it gets the least throughput compares with TCP and GAIMD. Yung Yi et al [9] have developed a fair hop-by-hop congestion control algorithm with the MAC constraint being imposed in the form of a channel access time constraint, using an optimization-based framework. In the absence of delay, they have shown that their algorithm is globally stable using a Lyapunov-function-based approach. Next, in the presence of delay, they have shown that the hop-by-hop control algorithm has the property of spatial spreading. Also they have derived bounds on the “peak load” at a node, both with hop-by-hop control, as well as with end-to-end control, show that significant gains are to be had with the hop-by-hop scheme, and validate the analytical results with simulation. Umut Akyol et al [10] have studied the problem of jointly performing scheduling and congestion control in mobile adhoc networks so that network queues remain bounded and the resulting flow rates satisfy an associated network utility maximization problem. They have defined a specific network utility maximization problem which is appropriate for mobile adhoc networks. They have described a wireless Greedy Primal Dual (wGPD) algorithm for combined congestion control and scheduling that aims to solve this problem. They have shown how the wGPD algorithm and its associated signaling can be implemented in practice with minimal disruption to existing wireless protocols. S. Karunakaran et al [11] have

presented a Cluster Based Congestion Control (CBCC) protocol that consists of scalable and distributed cluster-based mechanisms for supporting congestion control in mobile ad hoc networks. The distinctive feature of their approach is that it is based on the self-organization of the network into clusters. The clusters autonomously and proactively monitor congestion within its localized scope.

3 Problem Identification and Proposed Protocol Overview

3.1 Problem Identification

As explained in section 2, congestion adaptive routing has been examined in several studies. Estimating or reviewing the level of activity in the intermediate nodes using load or delay measurement, is the common approach in all the studies mentioned. The favorable path is established based upon the collected information, which helps in avoiding the existing and developing congested nodes. The performance of routing protocols is affected by the service type of the traffic carried by the intermediate nodes. But no research has stated this so far.

Before presenting themselves as aspirant to route traffic to the destination, the MANETs do not take the status of the queues into account, for the route discovery process. Because of this, the newly arriving traffic face long delays, packet drops, and fail to be transmit ahead of the already queuing traffic.

The mobile ad hoc networks performances are subjective to the congestion problem. A routing algorithm and a flow control scheme, includes the congestion control scheme. Enhanced performance and better congestion control can be achieved only by considering the routing and the flow control together. This was not done in earlier researches [12].

3.2 Protocol Overview

In this paper, we propose to design and develop an agent based congestion control technique. In our technique, the information about network congestion is collected and distributed by mobile agents (MA). Each node has a routing table that stores routing information for every destination. MA starts from every node and moves to an adjacent node at every time. The MA updates the routing table of the node it is visiting.

In our technique, the node is classified in one of the four categories depending on whether the traffic belongs to background, best effort, video or voice AC respectively. Then MA estimates the total congestion metric by calculating the queue length and the channel contention and it is applied to the routing protocol to select the minimum congested route.

4 Agent Based Congestion Control

4.1 EDCA Mechanism of 802.11e

The Hybrid Coordination Function (HCF) which has been sketched by 802.11e labels two new MAC methods. The PCF and DCF modes have been replaced with HCF controlled channel access (HCCA), and enhanced distributed channel access (EDCA) which provides distributed access supplying service differentiation [13].

An extended version of the legacy DCF mechanism is EDCA. Access Categories (AC) or traffic priority classes like voice, video, best effort and background are defined by EDCA [14]. The access categories prioritize themselves from AC3 to AC0. In general, best effort and background traffic are maintained by AC1 and AC0 and real-time applications like voice or video transmission are maintained by AC2 and AC3 [15]. For the purpose of service differentiation, many MAC constraints vary with priority level chosen for each AC.

For the implementation of the EDCA contention algorithm the four transmission queues are applied with each AC being communicated with the others. The minimum idle delay before contention (AIFS), the Contention Windows (CW_{min} and CW_{max}), and the Transmission opportunity limit (TXOP) are the various parameters described here. The default values of each parameter are listed in Table 1.

Table 1. IEEE 802.11e EDCA MAC System Parameters

Access Category	AIFSN	CW_{min}	CW_{max}	Queue length	Max. retry limit
AC3	2	7	15	25	8
AC2	2	15	31	25	8
AC1	3	31	1023	25	4
AC0	7	31	1023	25	4

In the MAC layer, voice traffic is conveyed through AC3 and the video traffic is conveyed through AC2 in accordance with 802.11e EDCA standard. The AC class differentiation in EDCA is very much useful in providing services to the traffic. Superior servicing is done for high-priority traffic and not much importance is given for low-priority traffic. The contention parameters of EDCA are not able to adapt to the network conditions, in spite of the delay sensitivity of real-time traffic taken into account. This leads to limitations in the QoS improvement [16].

ACs pause for diverse values of Arbitration Interframe Space (AIFS) and AIFSi is computed by,

$$AIFSi = SIFS + AIFSNi \times SlotTime$$

where AIFSi is a positive integer which is greater than one, AIFSNi is the AC-specific AIFS number; SIFS and Slot Time are dependent on physical layer [14]. If the values of the subsequent parameters are small, the channel access delay will become less for the AC which leads the higher priority to approach the medium.

When a particular QoS station (QSTA) has the concession to begin transmissions, then the TXOP is expressed as the time interval in IEEE 802.11e. The initiation of the TXOP and the multiple frame transmission within an EDCA TXOP are the nodes approved by TXOP. The former occurs only when the EDCA rules allow entry to the medium. And the later occurs when an EDCA Function (EDCAF) holds the concession to contact the medium after completing a frame exchange sequence. The period of TXOP values are herewith in the EDCA parameter engraved in beacon frames. A STA is allowed to transmit multiple MAC protocol data units (MPDUs) from the same AC with a SIFS time interval between an ACK and the succeeding frame transmission. A single MPDU may be forwarded for each TXOP if the TXOP limits the value of 0 [17].

4.2 Mobile Agent (MA)

A Mobile Agent (MA) starts from every node and moves to an adjacent node at every time. A node visited next is selected at the equivalent probability. The MA brings its own history of movement and updates the routing table of the node it is visiting.

Each MA has its own history which consists of its source node S, the current time Tc, the number of hops NH from the starting node, the adjacent node AN that the MA has last visited and the number of multiple packets NP on AN at Tc. When an MA visits a node, it puts the information (S, Tc, NH, AN, NP) in the routing table of that node.

Each node has a routing table that stores k fresh routing information records from itself to every node S: [S, {(Tc1, NH1, AN1, NP1) (Tcm, NHm, ANm, NPm)}], where Tc1 > Tc2 >..... > Tm. We call m the number of entries. For each i (1 ≤ i ≤ m), Tci is a time of visiting the adjacent node ANi, NHi is the number of hops and NPi is the number of MAs on ANi. When MA with the history (S, Tc, NH, AN, NP) visits a node N, the routing information on that node

[S, {(Tc1, NH1, AN1, NP1) (Tcm, NHm, ANm, NPm)}] is updated to

[S; {(Tc, NH; AN, NP), (Tc1, NH1, AN1, NP1)..... (Tcm-1, NHm-1, ANm-1, NPm-1)}].

4.3 Queue Length Estimation

The traffic rate within the network has to be determined to find the level of congestion. The traffic rate is significantly affected by

- the number of new incoming flows
- the number of existing flows
- the density of the nodes in the network
- Communication abilities of nodes

Our goal is to acquire macroscopic network statistics using a heuristic approach. We compute the traffic rate as follows: Let the value L_o represent the offered load at the queue of node i and it is defined as

$$L_{oi} = \frac{AR_i}{SR_i} \tag{1}$$

where AR_i is the aggregate arrival rate of the packets produced and forwarded at node i while SR_i is the service rate at node i, i.e., SR_i= 1/T where T is the computed exponentially weighted moving average of the packets' waiting time at the head of the service queue. The distribution of the queue length PR (Q_i) (essentially this is the probability that there are Q_i packets in the queue) at the node is computed as

$$PR (Q_i) = (1 - L_{oi}) L_{oi}^i \tag{2}$$

For N distinct queues, the joint distribution is the product

$$PR (Q_{11}, Q_{12} \dots Q_{1N}) = \prod_{i=1}^N (1 - L_{oi}) L_{oi}^{li} \tag{3}$$

4.4 Channel Contention Estimation

In this network, we consider IEEE 802.11 MAC with the distributed coordination function (DCF). It has the packet sequence as request-to-send (RTS), clear-to-send (CTS), data and acknowledgement (ACK). The amount of time between the receipt of one packet and the transmission of the next is called a short inter frame space (SIFS). Then the channel occupation due to MAC contention will be

$$C_{OCC} = t_{RTS} + t_{CTS} + 3t_{SIFS} + t_{acc} \tag{4}$$

Where t_{RTS} and t_{CTS} are the time consumed on RTS and CTS, respectively and t_{SIFS} is the SIFS period. t_{acc} is the time taken due to access contention.

The channel occupation is mainly dependent upon the medium access contention, and the number of packet collisions. That is, C_{occ} is strongly related to the congestion around a given node.

C_{occ} can become relatively large if congestion is incurred and not controlled, and it can dramatically decrease the capacity of a congested link.

4.5 Total Congestion Metric

The Total Congestion Metric (TCM) can be estimated from the obtained queue length and the channel contention.

$$TCM = PR(Q_i) + C_{occ} \tag{5}$$

5 Agent Based Congestion Control Routing

The agent based congestion routing can be explained from the following figure:

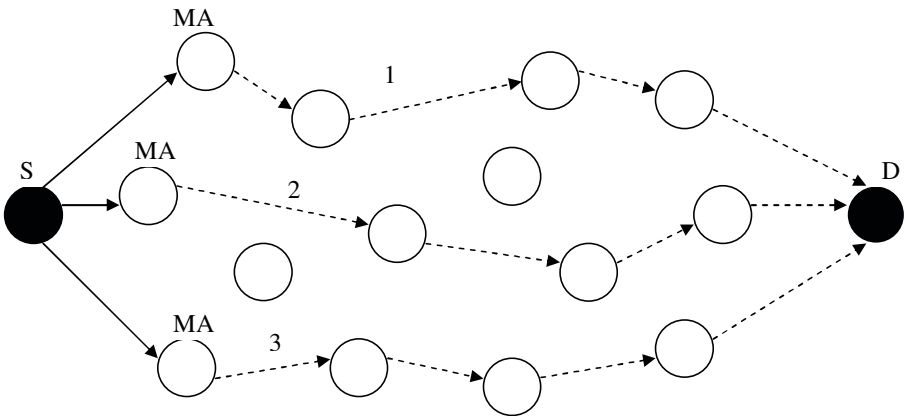


Fig. 1. Agent Based Congestion Routing

- Step 1: The source S checks the number of available one hop neighbors and clones the Mobile Agent (MA) to that neighbors.
- Step 2: The Mobile Agent selects the shortest path of the route to move towards the destination D as given in the figure 1 such as P1, P2 and P3.
- Step 3: The MA1 moves towards the destination D in a hop-by-hop manner in the path P1 and MA2 in P2 and MA3 in P3 respectively.
- Step 4: Then the MA1 calculates the TCM1 of that path P1 and similarly MA2 calculates the TCM2 of P2 and MA3 calculates the TCM3 of P3.
- Step 5: Now the destination D sends the total congestion metrics TCM1, TCM2 and TCM3 of the paths P1, P2 and P3 respectively to the source.
- Step 6: Now the source selects path using min (TCM1, TCM2, and TCM3) and sends the data through the corresponding path which has the minimum congestion.

6 Simulation Results

6.1 Simulation Model and Parameters

We use NS2 [18] to simulate our proposed technique. In the simulation, the channel capacity of mobile hosts is set to the same value: 11Mbps. In the simulation, mobile nodes move in a 1000 meter x 1000 meter region for 50 seconds simulation time. Initial locations and movements of the nodes are obtained using the random waypoint (RWP) model of NS2. It is assumed that each node moves independently with the same average speed. All nodes have the same transmission range of 250 meters. The node speed is 5 m/s. and pause time is 5 seconds. In the simulation, for class1 traffic video is used and for class2 and Class3, CBR and FTP are used respectively.

The simulation settings and parameters are summarized in table 2.

Table 2. Simulation Settings

No. of Nodes	50
Area Size	1000 X 1000
Mac	802.11e
Radio Range	250m
Simulation Time	50 sec
Routing Protocol	AODV
Traffic Source	CBR and Video
Video Trace	JurassikH263-256k
Packet Size	512
Mobility Model	Random Way Point
Speed	5m/s
Pause time	5 sec
MSDU	2132
Rate	250kb,500kb,.....1000Kb
No. of Flows	2,4,6,8 and 10

6.2 Performance Metrics

We compare the performance our Agent Based Congestion Control (ABCC) technique with the Hop by Hop algorithm [9]. The performance is evaluated mainly, according to the following metrics.

Packet Delivery Fraction: It is the ratio of the number of packets received successfully and the total number of packets sent.

Throughput: It is the number of packets received successfully.

Average end-to-end delay: The end-to-end-delay is averaged over all surviving data packets from the sources to the destinations.

6.3 Results

A. Effect of Varying Rates

In the initial experiment, we measure the performance of the proposed technique by varying the rate as 250, 500, 750 and 1000Kb.

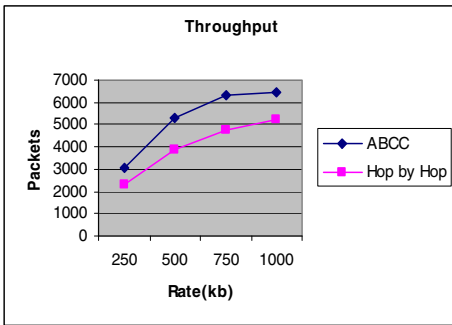


Fig. 2. Rate Vs Throughput

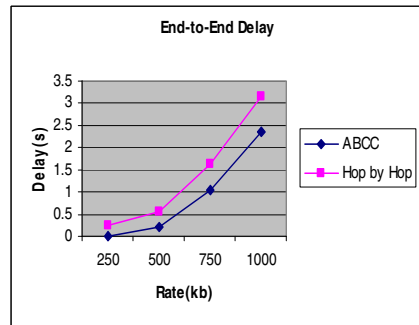


Fig. 4. Rate Vs End-to-End Delay

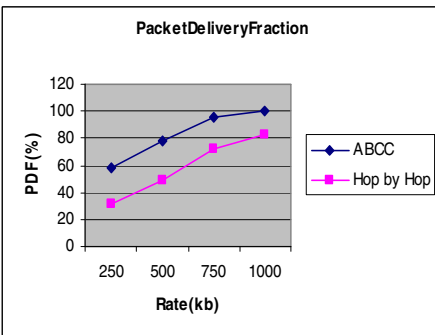


Fig. 3. Rate Vs Packet Delivery Fraction

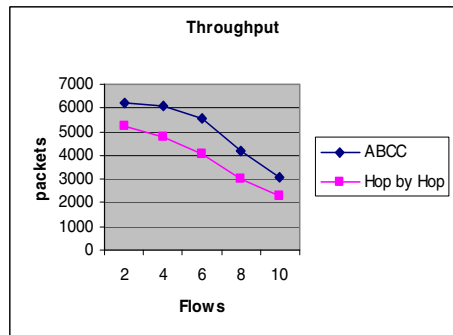


Fig. 5. Flows Vs Throughput

Figure 2 gives the throughput of the proposed technique when the rate is increased. As we can see from the figure, the throughput is more in the case of ABCC when compared to the Hop by Hop algorithm.

From Figure 3, we can see that the packet delivery fraction for ABCC is more, when compared to the Hop by Hop algorithm.

From Figure 4, we can see that the average end-to-end delay of the proposed ABCC technique is less when compared to the Hop by Hop algorithm.

B. Effect of varying Flows

In the next experiment, we compare our proposed technique by varying the number of flows as 2, 4, 6, 8 and 10.

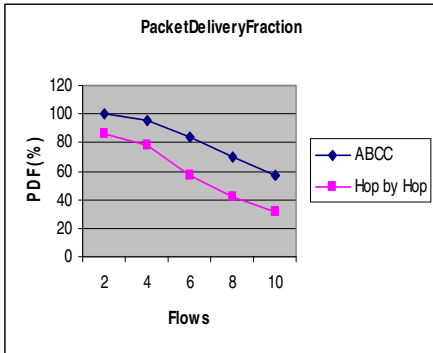


Fig. 6. Flows Vs Packet Delivery Fraction

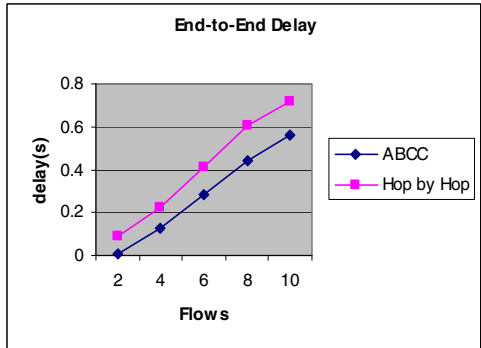


Fig. 7. Flows Vs End-to-End Delay

Figure 5 gives the throughput of the proposed technique when the flow is increased. As we can see from the figure, the throughput is more in the case of ABCC when compared to the Hop by Hop algorithm.

From Figure 6, we can see that the packet delivery fraction for ABCC is more, when compared to the Hop by Hop algorithm.

From Figure 7, we can see that the average end-to-end delay of the proposed ABCC technique is less when compared to the Hop by Hop algorithm.

7 Conclusion

In this paper, we have designed and developed an agent based congestion control technique. In our technique, the information about network congestion is collected and distributed by mobile agents (MA). A mobile agent starts from every node and moves to an adjacent node at every time. A node visited next is selected at the equivalent probability. The MA brings its own history of movement and updates the routing table of the node it is visiting. The MA updates the routing table of the node it is visiting. In this technique, the node is classified in one of the four categories depending on whether the traffic belongs to background, best effort, video or voice AC respectively. Then MA estimates the queue length of the various traffic classes and the channel contention of each path. Then this total congestion metric is applied to the routing

protocol to select the minimum congested route in the network. By simulation results, we have shown that our proposed technique attains high delivery ratio and throughput with reduced delay when compared with the existing technique.

References

1. Santhosh baboo, S., Narasimhan, B.: A Hop-by-Hop Congestion-Aware Routing Protocol for Heterogeneous Mobile Ad-hoc Networks. *International Journal of Computer Science and Information Security* (2009)
2. Chen, X., Jones, H.M., Jayalath, A.D.S.: Congestion-Aware Routing Protocol for Mobile Ad Hoc Networks. In: *IEEE 66th Conference in Vehicular Technology* (2007)
3. Lochert, C., Scheuermann, B., Mauve, M.: A Survey on Congestion Control for Mobile Ad-Hoc Networks. In: *Wireless Communications and Mobile Computing, InterScience* (2007)
4. http://en.wikipedia.org/wiki/Congestion_control
5. <http://www.linktionary.com/c/congestion.html>
6. Tran, D.A., Raghavendra, H.: Congestion Adaptive Routing in Mobile Ad Hoc Networks. *IEEE Transactions on Parallel and Distributed Systems* (November 2006)
7. Lien, Y.-N., Hsiao, H.-C.: A New TCP Congestion Control Mechanism over Wireless Ad Hoc Networks by Router-Assisted Approach. In: *27th IEEE International Conference on Distributed Computing Systems Workshops* (2007)
8. Sun, W., Wen, T., Guo, Q.: A Performance Comparison of Equation-Based and GAIMD Congestion Control in Mobile Ad Hoc Networks. In: *International Conference on Computer Science and Software Engineering* (2008)
9. Yi, Y., Shakkottai, S.: Hop-by-Hop Congestion Control Over a Wireless Multi-Hop Network. *IEEE/ACM Transactions on Networking* (February 2007)
10. Akyol, U., Andrews, M., Gupta, P., Hobby, J., Saniee, I., Stolyar, A.: Joint Scheduling and Congestion Control in Mobile Ad-Hoc Networks. In: *Proceedings of IEEE INFOCOM* (2008)
11. Karunakaran, S., Thangaraj, P.: A Cluster Based Congestion Control Protocol for Mobile Adhoc Networks. *International Journal of Information Technology and Knowledge Management* 2(2), 471–474 (2010)
12. Malika, B., Mustapha, L., Abdelaziz, M., Nordine, T., Mehammed, D., Rachida, A.: Intelligent Routing and Flow Control In MANETs. *Journal of Computing and Information Technology*, doi:10.2498/cit.1001470
13. Li, J., Li, Z., Mohapatra, P.: APHD: End-to-End Delay Assurance in 802.11e Based MANETs. In: *3rd Annual International Conference, Mobile and Ubiquitous Systems – Workshops*, pp. 1–8 (2006)
14. Lee, J.F., Liao, W., Chen, M.C.: A Differentiated Service Model for Enhanced Distributed Channel Access (EDCA) of IEEE 802.11e WLANs. In: *Proc. IEEE Globecom* (2005)
15. Ksentini, A., Naimi, M., Gueroui, A.: Toward an Improvement of H.264 Video Transmission over IEEE 802.11e through a Cross-Layer Architecture. *IEEE Communications Magazine* (January 2006)
16. Wu, Y.j., Chiu, J.-h., Sheu, T.-l.: A Modified EDCA with Dynamic Contention Control for Real-Time Traffic in Multi-hop Ad Hoc Networks. *Journal of Information Science And Engineering* 24, 1065–1079 (2008)
17. Flaithearta, P.O., Melvin, H.: 802.11e EDCA Parameter Optimization Based on Synchronized Time. In: *MESAQIN 2009* (2009)
18. Network Simulator, <http://www.isi.edu/nsnam/ns>

A Hybrid Algorithm for Satellite Image Classification

Samiksha Goel¹, Arpita Sharma², and V.K. Panchal³

¹ Department of Computer Science, Delhi University, India
Samiksha.goel@gmail.com

² Department of Computer Science, DDU College, Delhi University, India
asharma@ddu.du.ac.in

³ Defence Terrain Research Laboratory, DRDO, Delhi, India
vkpans@ieee.com

Abstract. Remote sensing is the most relevant science that permits us to acquire information about the surface of the land, without having actual contact with the area being observed. Amongst the multiple uses of remote sensing, one of the most important has been its use in solving the problem of land cover mapping. Multi spectral classification of remotely sensed data has been widely used to generate thematic Land-Use/Land-Cover maps. Two of the extensively used algorithms for image classification are Self Organizing Feature Maps (SOFM) and Ant Colony Optimization. Although both are bio-inspired optimization techniques, however combining them is a challenging task, especially in the field of remote sensing. In this paper, we have used a Self Organizing Ant Algorithm for Classification of remotely sensed data. Also, we have suggested a new reinforcement factor for the pheromone updation. A test of algorithm is conducted by classifying a high resolution, multi-spectral satellite image of Alwar Region. Results obtained are encouraging.

Keywords: Ant Colony Optimization, SOFM, Image classification.

1 Introduction

The methods employing remote sensing techniques for extraction of urban land use information and subsequent analysis and modeling have evolved from a very basic visual interpretation into a complicated family. In last two decades, many advanced classification approaches, such as Artificial Neural Networks, Fuzzy sets, Expert Systems and Genetic Algorithms have been widely used for image classification [1,2]. In the recent past a new range of computational algorithms known as Swarm Intelligence Algorithms, which are inspired from the behaviour of social insects, have gained the attention of researchers for classification of remotely sensed data[3,4]. Two widely used swarm intelligence algorithms are Ant Colony Optimization (ACO) [5] and Particle Swarm Optimization (PSO) [6]. Hybrid approaches using these two and other already existing techniques are being experimented with currently[7,8]. Self Organizing Ant Algorithm [7] is one such hybrid algorithm which merges the concepts of Kohonen Self organizing feature map[9] and Ant Algorithm[10,11]. A

modified version of this algorithm is used in this paper for the satellite image classification.

The remaining paper is organized in as follows. A brief description of the algorithm used is given in section 2. Section 3 comprises of data acquisition and preparation. Experiments and result are discussed in section 4. Conclusion and future work makes the final section 5 of the paper.

2 Methodology

The algorithm presented in this paper combines the features of two extensively used algorithms, i.e., Self Organizing Feature Maps (SOFM) and Ant Colony Optimization. The basic idea here is to use stigmergy for clustering and classification. Usual ant clustering algorithm place data as objects in the grid, ants move around, and then, via some natural inspiration and a great deal of heuristics, they manage to cluster them according to proximity. Self-Organizing Ant Algorithm, on the other hand, makes each data item an ant. Pheromones are also vectorial in nature, in the same dimension as data.

The main idea is to assign a random vector (having the same dimension and range as the training set) to each cell of the grid that makes the environment. Then assign each input sample vector (training data) to an ant and put them randomly into a cell on the grid. The ants then move around in the grid and simultaneously change the grid vector to bring them closer towards their own vector. The updation of the grid vector of the cell, where the ant has moved is considered as pheromone updation operation. The updation depends upon the ant vector and the centroid vector of the zone centered at the cell where ant has moved. Every ant tends to move towards those zones in the grid which have vectors more similar to the ant vector[7]. Eventually, ants with similar data items will be closer together in the grid and the grid itself will contain similar vector to those stored in the ants on top of them. The grid, then can be used as a classification tool while ants will be grouped in clusters of similar individuals.

The algorithm contains the following steps.

- Grid vectors are initialized randomly.
- Assign training vectors of each class to ants and place them randomly on the grid.
- Set the loop for N iterations
- For each ant a placed at cell i decide the position of the next cell j , where ant has to move, as follows:
 - Consider a neighbourhood N_i^t of radius nr of cell i at time t . For each cell k within this neighbourhood N_i^t
 - ❖ Consider a neighbourhood N_k of radius cr and calculate the centroid vector CTR_k of all the cells in the neighbourhood N_k . (CTR_k is a vector where each component takes the arithmetic mean of the corresponding component of vectors of the cells falling in the neighbourhood N_k)

❖ Compute σ_{ik} which is the Euclidean distance between the vector (V_i) associated to the cell i and the centroid vector CTR_k , both vectors have n components.

$$\sigma_{ik} = \sqrt{\sum_{v=1}^n (V_i(v) - CTR_k(v))^2} \tag{1}$$

❖ Also compute Ant system pheromone weighing function $W(\sigma_{ik})$ and the transition probability P_{ik}

$$W(\sigma_{ik}) = \left(1 + \frac{\delta_a}{1 + \sigma_{ik} \cdot \delta_a} \right)^{\beta_k} \tag{2}$$

Where, δ_a defines the ability of ant a to sense the pheromone. And β_k defines the pheromone concentration at cell k . If β_k is low, pheromone concentration doesn't affect the choice.

$$P_{ik} = \frac{W(\sigma_{ik})}{\sum_{u \in N_i^t} W(\sigma_{iu})} \tag{3}$$

❖ Now generate a random number q between $[0,1]$. For selecting the cell j , where ant will move, pseudo-random proportional rule of Ant Colony System is used. Which is as follows-

If ($q < q_0$) (4)
 $j = \arg \max W(\sigma_{ik})$

else
 make roulette-wheel selection considering P_{ik} as probability of every feasible neighbour k . Here q_0 is a constant between 0 and 1.

➤ This cell is the cell j at which the ant will move.

- After finding the appropriate cell j for each ant, the grid vector of each target cell will be updated. The updation uses the reinforcement factor R (which is based on centroid vector CTR_j of j_{th} cell, ant vector a_k and learning rate α) and the original grid vector for the target cell.

$$R = \alpha \cdot (1 - \bar{D}(a_k, CTR_j)) \tag{5}$$

where

$$\bar{D} = \frac{\sqrt{\sum_{v=1}^n (a_k(v) - CTR_j(v))^2}}{n} \tag{6}$$

$$V_j^t(v) = V_j^{t-1}(v) + R \cdot (a_k(v) - V_j^{t-1}(v)) \tag{7}$$

$$\forall v = 1..n$$

- This process continues for N iterations.
 The reinforcement function R given by equation (5) in [7] is not suitable for satellite images. A new reinforcement function for pheromone updation is suggested by us in section 4.

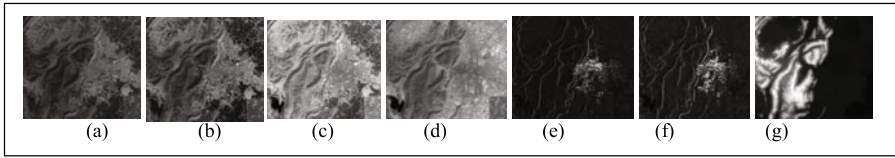


Fig. 1 (a-g). Seven band Grey scale images

3 Data Acquisition and Preparation

The satellite image of alwar region is obtained from the Indian Remote Sensing satellite's (IRS-P6) optical band image set (Figure 1) *i.e.* Green (G) (figure 1.a), Red (R) (figure 1.b), Near-Infrared (N) (figure 1.c) and Middle-Infrared (M) (figure 1.d) bands. The ground resolution of these images is 23.5m and is taken from LISS (Linear Imaging Self Scanning Sensor)-III, sensor.

The land cover classification with independent attributes set x consists of two sets of radarsat microwave images radarsat-1(r1) (figure 1.e) and radarsat-2(r2) (figure 1.f) and digital elevation model (d) (figure 1.g) data. Major land-use types of the area are vegetation, urban, water, rocky and barren. The size of the image is of size 472 X 546 and it contains 2,57,712 pixels.

The training data is provided by an expert using ERDAS software. Also, it has been ensured that each sample of training data was unique and should only belong to one class. The training data was provided for each of the 5 major land use types- water, vegetation, urban, barren and rocky, in the excel sheet format. The number of pixels in each class provided by the expert as training set is provided in table 1. Also for the purpose of our experiments, we have first converted these images as excel sheets, which is prepared in the same format as the training data is provided.

Table 1. Number of training vectors in each class

Class Name	Number of training Pixel
Water	206
Vegetation	329
Barren	417
Urban	288
Rocky	191

4 Experiments

For experimental purpose, we have taken the 10-by-10 grid for output layer. The classified image is shown in figure 2. In order to apply the algorithm to satellite images it was observed that the reinforcement function R (equation 5) is not suitable. So we have changed the Reinforcement factor for pheromone updation of the grid vector as follows-

$$R = \alpha (D(a_k, CTR_i) / (1 + D(a_k, CTR_i)))^2 \quad (8)$$

Here α is the initial learning rate and is taken as 0.5. Rectangular neighbourhood is chosen for the simplicity. Various experiments are conducted to examine the impact of neighbourhood radius nr and centroid radius cr . It has been found that the results enhanced when both are set to value 1. The algorithm uses two more parameters which are associated with the pheromone concentration(β) and the ability of ant to sense the pheromone (δ). In order to get suitable values for β and δ many runs were conducted on a small portion of the complete image. Images of the small classified portion are shown in Figures 3(a-f). The results obtained with values $\beta=8$ and $\delta=0.3$ are relatively better. Parameter α_0 is taken as 0.9.

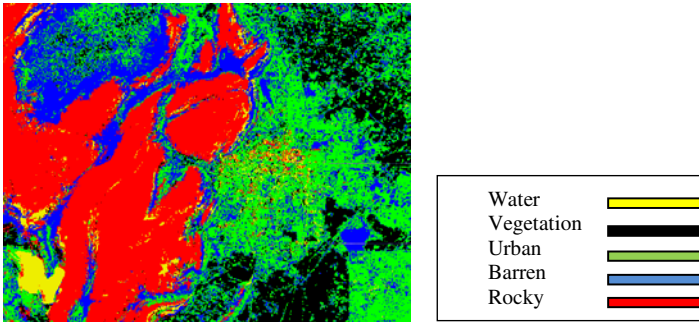


Fig. 2. Classified image using Self Organizing Ants

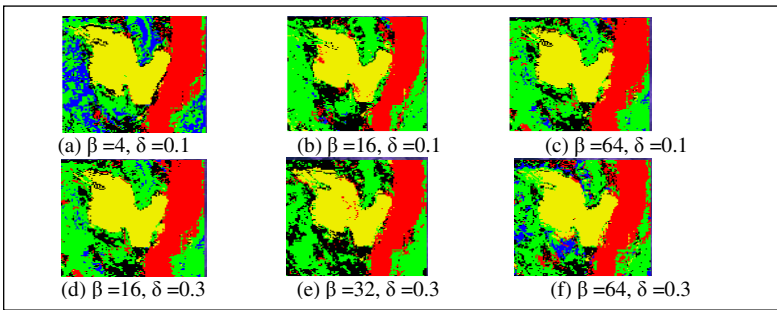


Fig. 3 (a-f). Classified images obtained when different combinations of β and δ are used in self organizing ant algorithm (shows results for left bottom corner of the image)

5 Results and Discussion

The error matrix, Kappa Coefficients and percentage accuracies of each class are shown in Table 2 and Figure 4 respectively. From figure 4, it can be seen that the three classes, Vegetation, Rocky and Water are predicted with high accuracy whereas accuracy of the prediction of the remaining two classes, though not very high but still more than 60%.

Table 2. Error Matrix of Self Organizing Ant with Kappa Coefficient: 0.70755

	Vegetation	Urban	Rocky	Water	Barren	Total
Vegetation	144	22	5	5	7	183
Urban	0	121	21	0	47	189
Rocky	6	0	161	0	5	172
Water	0	1	2	65	1	69
Barren	0	46	11	0	110	167
Total	150	190	200	70	170	780

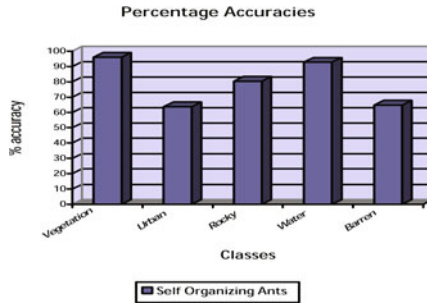


Fig. 4. Bar graph for percentage accuracy for each class

In order to see the effect of grid size on the accuracy of the prediction, we have test the program with varying grid sizes and observed that the grid size doesn't play a major role in deciding the accuracy of the classification. Kappa coefficient in general, varies between 0.60 and 0.70. The graph in figure 5(a), represents the variation of Kappa coefficients calculated by taking different grid sizes. Also, in figure 5(b), the graph shows the percentage accuracies for each class for different grid sizes.

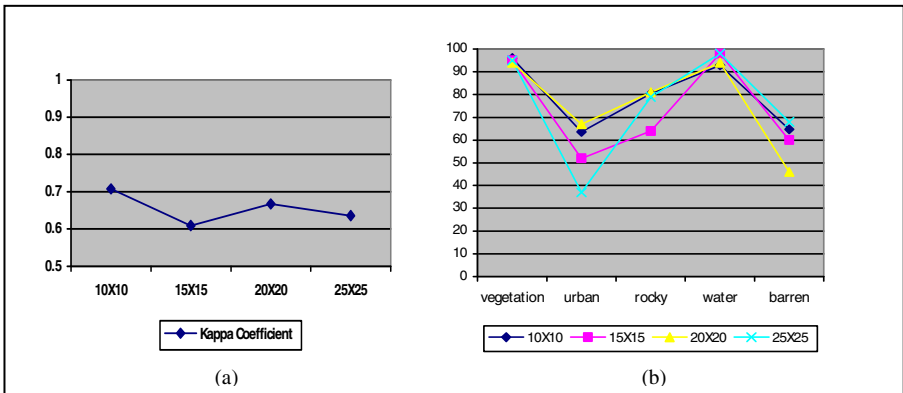


Fig. 5(a). Graph showing different kappa coefficients obtained by taking different grid sizes, and **(b)** : Class-wise accuracies obtained with different grid sizes

6 Conclusion and Future Work

A hybrid algorithm combining the concepts of Self Organizing map and Ant Colony Optimization is used in this paper to classify the satellite image of Alwar region. The initial results obtained are promising. There is a scope of further improvement in the results. In this paper, only the vector associated with the cell where an ant has moved, is modified to make it closer to the ant vector whereas the vector associated with the cells in the neighbourhood of the cell should also be modified as centroid vector is the mean of the vectors of the neighbouring cells, and decision of moving the ant to a cell is based on the values of W or P_{ij} which in turn depend on σ . σ itself is calculated as Euclidean distance between grid vector of the cell where ant is placed and centroid vector. Besides, few other functions for the reinforcement factor R , may also be tested to improve the results further.

References

1. Lu, D., Weng, Q.: A Survey of Image Classification Methods and Techniques for improving classification performance. *International Journal of Remote Sensing* 28(5), 823–870 (2007)
2. Mather, P., Tso, B.: *Classification Methods for Remotely Sensed Data*, 2nd edn. CRC Press, Boca Raton (2009)
3. Khedam, Outemzabet, R., Tazaoui, N., Belhadj-Aissa, Y.: A Unsupervised Multispectral Image Classification using Artificial Ants. In: *Information and Communication Technologies, ICTTA 2006*, Fac. of Electron. and Comput. Sci., Univ. of Sci. and Technol. Houari Boumediene, Algiers, pp. 349–354 (2006)
4. Omkar, Manoj, S.N., Mudigere, K.M., Dipti Muley, D.: Urban Satellite Image Classification using Biologically Inspired Techniques. In: *IEEE International Symposium on Industrial Electronics, ISIE 2007*, June 4-7, pp. 1767–1772 (2007)
5. Colomi, A., Dorigo, M., Maniezzo, V.: Distributed Optimization by Ant Colonies. In: *actes de la première conférence européenne sur la vie artificielle*, Paris, France, pp. 134–142. Elsevier Publishing, Amsterdam (1991)
6. Kennedy, J., Eberhart, R.: *Proc. IEEE Int'l. Conf. on Neural Networks*, Perth, Australia, vol. IV, pp. 1942–1948. IEEE Service Center, Piscataway (1995)
7. Fernandes, C., Mora, A.M., Merelo, J.J., Ramos, V., Laredo, J.L.J.: Kohonants: A self organizing ant algorithm for clustering and pattern classification. In: *Artificial Life XI 2008* (2008)
8. Chandramouli, K.: Particle Swarm Optimisation and Self Organising Maps based Image classifier. In: *Second International Workshop on Semantic Media Adaptation and Personalization*. IEEE, Los Alamitos (2007)
9. Kohonen, T.: *Self organizing maps*. Springer, New York (2001)
10. Dorigo, M., Maniezzo, V., Colomi, A.: Ant System: Optimization by a Colony of Cooperating Agents. *IEEE Transactions on Systems, Man, and Cybernetics–Part B* 26(1), 29–41 (1996)
11. Chialvo, D., Millonas, M.: How Swarms build cognitive maps. In: *The Biology and Technology of Intelligent Autonomous Agents*. NATO ASI Series, vol. 144, pp. 439–450 (1995)

An Image Authentication Technique in Frequency Domain Using Secure Hash Algorithm (FDSHA)

Amitava Nag¹, Debasish Biswas¹, Soumadip Ghosh¹, Sushanta Biswas²,
Debasree Sarkar², and Partha Pratim Sarkar²

¹ Academy of Technology, West Bengal University of Technology, Hoogly – 721212, India
it_amitava@yahoo.co.in

² DETS, University of Kalyani, Kalyani, Nadia – 741 235, West Bengal, India
ppsarkar@klyuniv.ac.in

Abstract. The demand of security is getting higher in these days due to the development of computers and Internet. In this paper, we proposed a novel image authentication scheme in frequency domain using SHA-1 that allows two parties to exchange images while guaranteeing image authentication, integrity and non-repudiation from the image sender, over an unsecured channel. The main idea is that at sender end, an image hash of the authenticating image of size 160 bits is generated using SHA-1 and this 160 bits then encrypted by sender's public key to produce a signature. Then the signature is embedded into cover image in frequency domain along with the authenticating image and creates stego-image with a minimum amount of perceivable degradation to the "cover" media.

Keywords: Steganography, Frequency Domain, Discrete Cosine Transform (DCT), SHA-1, Information Hiding, Authentication.

1 Introduction

Since digital images and video are now widely distributed via the internet and various public channels, Image authentication technology is becoming increasingly important. Image authentication is a technique for inserting information into an image for identification and authentication. To cater this, data hiding [1, 6] algorithms are used to embed the secret information. Data hiding is an old but interesting technology. Steganography (Greek means "covered writing") is a branch of data hiding which represents a class of processes used to embed secret data inside various forms of ordinary media such as image, audio, video, or text with a minimum amount of perceivable degradation to the "cover" media and create a stego-media.

The Data hiding technology can be broadly classified into two categories: (1) spatial domain based and (2) frequency domain based approaches. In spatial domain approaches, the secret messages are embedded directly. On spatial domain, least significant bit (LSB) is the most commonly used type of insertion scheme used currently in digital steganography. However, the LSB insertion method is easy to be attacked. Possible frequency image transformations include the discrete Fourier transform (DFT) [11], discrete cosine transform (DCT) [8,10,13], and others [7]. The DCT is a

mathematical transformation that takes a signal and transforms it from spatial domain into frequency domain [8,10]. The DCT transformation is adopted in this paper.

Several methods are developed to find hash function [2-5]. Here SHA-I algorithm is used to generate unique 160 bits. Data hiding [6] in the image has become an important technique for image authentication and identification.

2 Proposed Algorithm

Spatial domain methods are less complex as no transform is used, but are not robust against attacks. In contrast, the algorithms in transform domain, such as DCT [8,10], DFT [11] and DWT [11, 12] have certain robustness. In this paper, we propose an information-hiding technique based on discrete cosine transform (DCT) of cover image for hiding a large amount of data with high security, a good invisibility and no loss of secret message. Moreover in this paper we use an authentication process by signature verification that is generated using SHA-1 which has strong potentiality in authentication checking of secret information.

The basic idea to hide authenticating message/image and a signature generated from authenticating message/image in the frequency domain is to alter the magnitude of all of the DCT coefficients of cover image. The 2-D DCT convert the image blocks from spatial domain to frequency domain. The schematic/ block diagram of the whole process is given in figure 1.

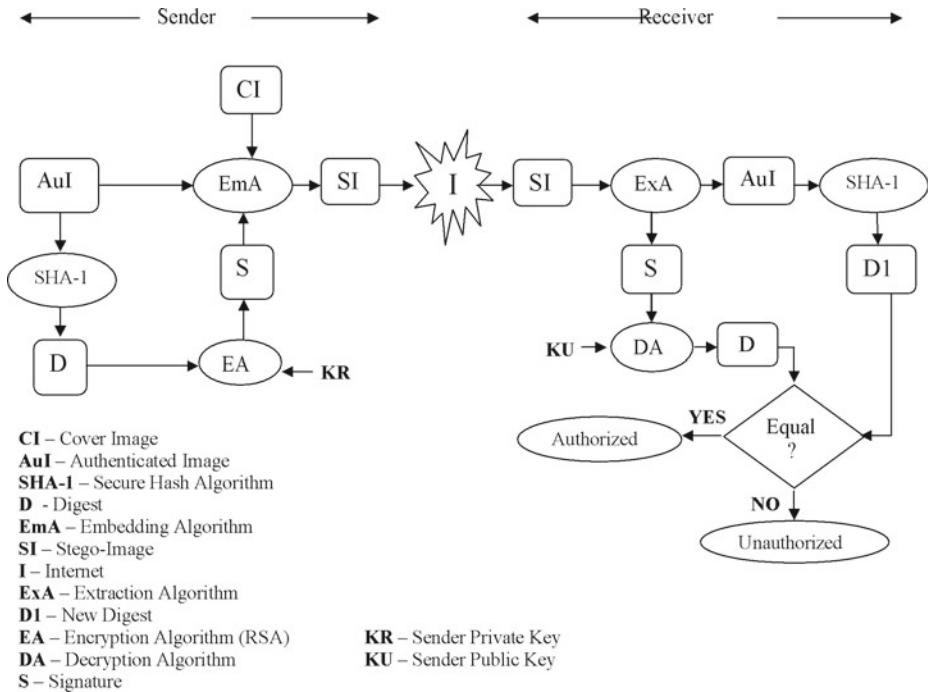


Fig. 1. Proposed Authentication Model

A. Discrete Cosine Transform

Let $I(x,y)$ denote an 8-bit grayscale cover-image with $x = 1,2,\dots,M_1$ and $y = 1,2,\dots,N_1$. This $M_1 \times N_1$ cover-image is divided into 8×8 blocks and two-dimensional (2-D) DCT is performed on each of $L = M_1 \times N_1 / 64$ blocks. The mathematical definition of DCT is:

Forward DCT:

$$F(u,v) = \frac{1}{4} C(u)C(v) \sum_{x=0}^7 \sum_{y=0}^7 f(x,y) \cos\left[\frac{\pi(2x+1)u}{16}\right] \cos\left[\frac{\pi(2y+1)v}{16}\right] \quad (1)$$

for $u = 0, \dots, 7$ and $v = 0, \dots, 7$

$$\text{where } C(k) = \begin{cases} 1/\sqrt{2} & \text{for } k = 0 \\ 1 & \text{otherwise} \end{cases}$$

Inverse DCT:

$$f(x,y) = \frac{1}{4} \sum_{u=0}^7 \sum_{v=0}^7 C(u)C(v) F(u,v) \cos\left[\frac{\pi(2x+1)u}{16}\right] \cos\left[\frac{\pi(2y+1)v}{16}\right] \quad (2)$$

for $x = 0, \dots, 7$ and $y = 0, \dots, 7$

B. Embedding Technique

Divide the carrier image into non overlapping blocks of size 8×8 and apply DCT on each of the blocks of the cover image f to obtain F using eqⁿ (1). Suppose S is the original 8-bit gray-level authenticating-image of size $M \times N$. It is denoted as

$$S = \{p_{ij} \mid 1 \leq i \leq M, 1 \leq j \leq N, p_{ij} \in \{0,1,2 \dots \dots 255\}\}$$

Now we convert S into an 1-D array S' as follows

$$S' = \{x_k \mid 1 \leq k \leq M \times N, x_k \in \{0,1,2 \dots \dots 255\}\}$$

Now each of individual pixels x_k in S' is made up of a string of bits of length 8 and form a 1-D binary string denoted as

$$B = \{b_k \mid 1 \leq k \leq 8 \times M \times N, b_k \in \{0,1\}\}$$

The least significant bit of all of the DCT coefficients inside 8×8 block is changed to a bit taken from each 8 bit block B from left to right. The method is as follows:

$$\begin{aligned} &\text{For } k=1 ; k \leq 1; k=k+1 \\ &\text{LSB}((F(u,v))_2) \leftarrow B(k) ; \end{aligned}$$

Where $B(k)$ is the k^{th} bit from left to right of a block B and $(F(u,v))_2$ is the DCT coefficient in binary form. Perform the inverse block DCT on F using eqⁿ (2) and obtain a new image f_1 which contains secret image.

Embedding Algorithm

Input: An $M_1 \times N_1$ cover image and an authenticating message/image.

Output: A stego-image.

1. Compute 160 bits SHA-1 digest (say D) of authenticating message/image (S) using SHA-1 algorithm.
2. Generate a 160 bits signature (S) from the 160 bits digest (D) obtained in step 1 by encrypting (RSA) the private key of sender.
3. Convert the 2-D authenticating image (AuI) of size $M \times N$ to the 1-D binary string of length $8 \times M \times N$.
4. Compute size of the bits stream obtained in step 3 in bits.
5. Divide the carrier image into non overlapping blocks of size 8×8 and apply DCT on each of the blocks of the cover image.
6. Repeat for each bit obtained in step 4
 - (a) Insert the bits into LSB position of each DCT coefficient of 1st 8×8 block found in step 4.
7. Repeat for each bit obtained in step 3
 - (a) Change the LSB of each DCT coefficient of each 8×8 block (excluding the first) found in step 4 to a bit taken from left (LSB) to right (MSB) from 1-D binary string obtained in step 2.
8. Repeat for each bit of the 160 bits signature (S) obtained in step 2
 - (a) Insert the bits into LSB position of each DCT coefficient
9. Apply inverse DCT using identical block size.
10. End.

Extraction Algorithm

Input: An $M_1 \times N_1$ Stego-image.

Output: Secret image.

1. Divide the stego-image into non overlapping blocks of size 8×8 and apply DCT on each of the blocks of the stego-image.
2. The size of the bit stream of 1-D binary string is extracted from 1st 8×8 DCT block by collecting the least significant bits of all of the DCT coefficients inside the 1st 8×8 block.
3. The least significant bits of all of the DCT coefficients inside 8×8 block (excluding the first) are collected and added to a 1-D array.
4. Repeat step 3 until the size of the 1-D binary string becomes equal to the size extracted in step 2.
5. Construct the authenticating image (AuI) from the bits extracted in step 3.
6. Extract least significant bits of the remaining DCT coefficients inside 8×8 block until 160 bits are extracted and generate a signature (say S) and decrypt it by the sender's public key to produce 160 bits digest (say D).
7. Again generate SHA - 1 (say D1) from extracted authenticating message/image (AuI) generated in 5 using SHA-1 algorithm.
8. Compare D (obtained in step 6) and D1 (obtained in step 7)

- (a) If D and $D1$ are equal – AuI is authorized and accept it otherwise AuI is unauthorized and reject it.
9. End.

3 Results and Discussion

In this section, some experiments are carried out to prove the efficiency of the proposed scheme. The proposed method has been simulated using the MATLAB 7 program on Windows XP platform. A set of 8-bit grayscale images of size 512×512 are used as the cover-image to form the stego-image (Fig 2a). The authenticating image 'Cameraman' (Fig 2b) of size 170×170 has been embedded to 'Lena' using FDSHA algorithm.

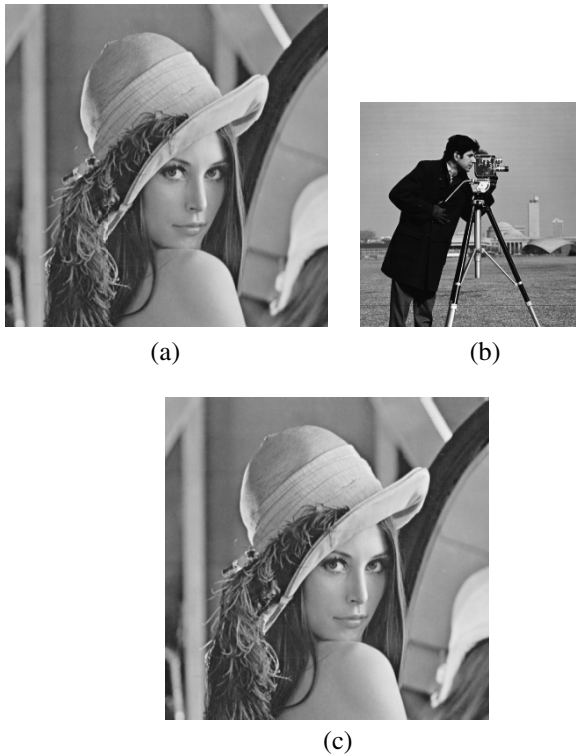


Fig. 2. (a) Cover-image, (b) Authenticating image and (c) Stego-image

The hiding capacity of authenticating messages/images is about 30×10^4 bits into a 512×512 cover (carrier) image. Here we are embedding a 8-bit grayscale image of size 170×170 into a 8-bit grayscale images of size 512×512 i.e. 231200 bits are embedded into a 512×512 carrier image. Table 1 exhibit the capacity and PSNR of four images. From table 1, it is observed that for all images, PSNR is greater than 50.

Table 1. Capacity and PSNR of four images

Images	Sise(in Kbyte)	Capacity(bits)	PSNR
Lena	256	231200	50.48
Baboon	256	231200	50.28
Airplane	256	231200	50.91
Boat	256	231200	50.36

Since the PSNR is high (for Lena 50.48 dB), one can barely distinguish the difference from the cover images shown in Fig. 2.

4 Conclusions

In this paper, the major attention is given on the secrecy as well as the authenticity of information. In this paper we presents an algorithm to insert large volume of messages/image data along with 160 bit signature that is generated from authenticating message/image using SHA – 1 inside the cover image for image authentication. The proposed technique is a frequency domain scheme developed expressly for oblivious authenticating image. Most of the works [1, 7] use embedding algorithm in time domain, but in our proposed method the embedding process is hidden under the transformation i.e. DCT and inverse DCT. These operations keep the images away from stealing, destroying from unintended users and hence the proposed method may be more robust against brute force attack.

References

- [1] Pfitzmann, B.: Information Hiding Terminology. In: Anderson, R. (ed.) IH 1996. LNCS, vol. 1174, pp. 347–350. Springer, Heidelberg (1996)
- [2] Rivest, R.: The MD4 Message Digest Algorithm 1320. MIT and RSA Data Security Inc. (April 1992)
- [3] National Institute of Standards and Technology, Fips 180, Federal Information Processing Standards, Secure Hash Standard (SHS) (April 1993)
- [4] Eastlake III, D., Jones, P.: US Secure Hash Algorithm-1(SHA-1) (September 2001)
- [5] Stalling, W.: Cryptography and network Security Principles and Practice, 4th edn. Prentice-Hall, Englewood Cliffs
- [6] Amin, P., Lue, N., Subbalakshmi, K.: Statistically secure digital image data hiding. In: IEEE Multimedia Signal Processing MMSP 2005, China, pp. 1–4 (October 2005)
- [7] Liu, L.: A Survey on Digital Watermarking Technologies, Technical Report, Stony Brook University, New York, USA (2005)
- [8] Coconu, C., Stoica, V., Ionescu, F., Profeta, D.: Distributed implementation of discrete cosine transform algorithm on a network of workstations. In: Proceedings of the International Workshop Trends & Recent Achievements in Information Technology, Romania, pp. 116–121 (2002)
- [9] 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)

- [10] Chu, R., You, X., Kong, X., Ba, X.: A DCT-based image steganographic method resisting statistical attacks. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), May 17-21, vol. 5, pp. V-953-6 (2004)
- [11] Kang, X., Huang, J., Shi, Y.Q., Lin, Y.: A DWT-DFT Composite Watermarking Scheme Robust to both affine Transformation and JPEG Compression. IEEE Transactions on Circuits and Systems for Video Technology 13(8) (August 2003)
- [12] Chen, P.-Y., Lin, H.-J.: A DWT Based Approach for Image Steganography. International Journal of Applied Science and Engineering 4(3), 275–290 (2006)
- [13] Nag, A., Biswas, S., Sarkar, D., Sarkar, P.P.: A novel technique for image steganography based on Block-DCT and Huffman Encoding. International Journal of Computer Science and Information Technology 2(3), 102–112 (2010)

A Robust and Fast Text Extraction in Images and Video Frames

Anubhav Kumar¹, Awanish Kr. Kaushik¹, R.L. Yadav¹, and Anuradha²

¹ Department of Electronics and Communication Engineering,
Galgotias College of Engineering and Technology, Greater NOIDA, India

² Department of Electronics and Communication Engineering,
Laxmi devi Institute of Engineering and Technology, Alwar, Rajsthan, India
rajput.anubhav@gmail.com, kaushik2feb@gmail.com

Abstract. Text detection in a color images is a very challenging problem. This paper gives an algorithm for detecting text in images. Experimental results on indoor, outdoor, captcha and Video frames images show that this method can detect text characters accurately. In this paper the proposed algorithm define the unite effect of the advantages of various previous approaches to find out the text, and focus on finding the text. Our experimental result on four different type images show that the technique based on line edge detection is reasonably better than the existing approach. This algorithm has 95.29% recall rate and average computed time is 3.645 second for English text. This is quicker than other existing methods and is robust to language, font- color and size.

Keywords: Line detection masks, Text detection, Text localization, Text extraction.

1 Introduction

Text detection in video and image has attracted researchers' attention for many years. As a result, hundreds of thousands of hours of archival videos are being stored and shared. Three types of text in images are: indoor -outdoor text which naturally occurs in the field of view of the camera and caption/graphics/artificial text which is artificially superimposed on the video at the time of editing and animation text which occurs in the field of view of the internet like captcha. The tough task in images is text extraction due to complicated background, ambiguous text character colors and different stroke specification.

There are two common methods are used to calculate the spatial connection which is based on edge based feature and connected component features of text. The area with a higher contrast between text and background focus upon Edge based method [2]. In this way, edges from letters are identified and merged. Connected component method [3] used a bottom-up approach by iteratively merge sets of connected pixels using a homogeneity criterion leading to the creation of flat-zones or Connected Components.

Our proposed method for image text extraction system (shown in Fig. 1) extract a text region from an image which can be broadly classified into three basic section:

(1)detection of the text region in the image, (2)localization of the region, and (3) extracted the output character image. Any possibility of text detection involves of text in the image is detected and the process of localization involves further enhancing the text regions by eliminating non-text regions. At last in text extraction process generates an output image with white text against a black background.

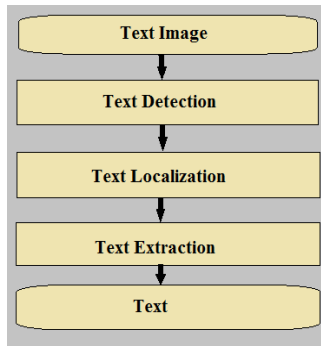


Fig. 1. Text image extraction flow chart

In this paper we focused on text extraction of four type's images with the help of line edge detector. The paper is organized as: Section 2 gives the proposed Algorithm and Section 3 gives the Experiment results is explained. Section 4 gives conclusions respectively.

2 Proposed Algorithm

In this section, the processing steps of the proposed text extraction are presented. Our aim is to build fast and robust text detection system which is able to handle still images with complex background. We can see from figure 1 that the proposed algorithm is mainly performed by three sections, which will be described below.

2.1 Text Detection

In our proposed method, images are next convolved with directional filters at different orientation masks for line edge detection in the horizontal (0° or 180°),vertical (90° or 270°) directions [1]. So it can be imagine that the next region have higher edge strength in the same directions. The line detection masks used are shown in Figure 2. Which enhance the find text edge, and then calculate the threshold. If the threshold of the detected edge set an appropriate value, than the other detected weak edge can be filtered.



Fig. 2. Line detection masks in Horizontal and Vertical directions

The basic steps of our algorithm are given below.

Step-1. Create line detection masks to detect edges at 0 or 180 and 90 or 270 orientation. So that we can find the directional edge maps which are used to represent the vertical and horizontal directions edge density and edge strength. In figure 3-d, the average edge image is shown. Figure 3-b, c shows the edge image in horizontal and vertical direction.

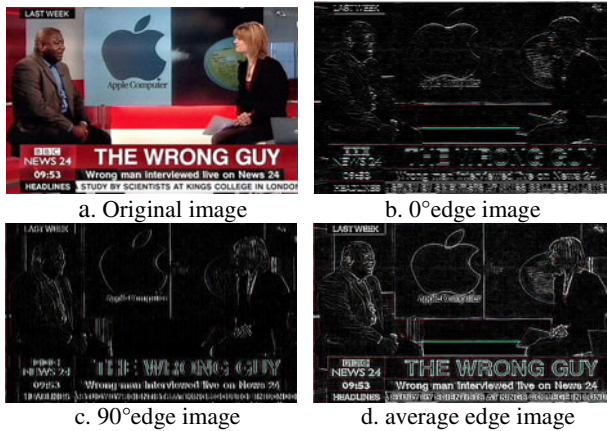


Fig. 3. Line Edge based Figure

Step-2. Convert edge image to binary image based on Otsu threshold.

Step-3. After taking Otsu threshold, the morphologically operation apply in the image and generally the morphological operations is used for binary images. Morphological result shown in figure 4.

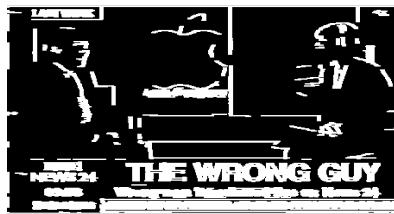


Fig. 4. Morphological result

2.2 Text Localization

In this section, the processing steps of the proposed text localization approach are presented.

Step 4. An analysis of horizontal and vertical projection files for the text region is placed. These profiles are mandatory histograms, in which every bin is a count of total pixel numbers in any existing row and column. The vertical and horizontal projection profiles for the sharpened edge image from Figure 4 are shown in Figure5 (a) and (b).

Step-5. Calculate the horizontal and vertical projection profiles of dilated image using histogram with an appropriate threshold value and Create refined image using multiplication of binary image and median filtering image.

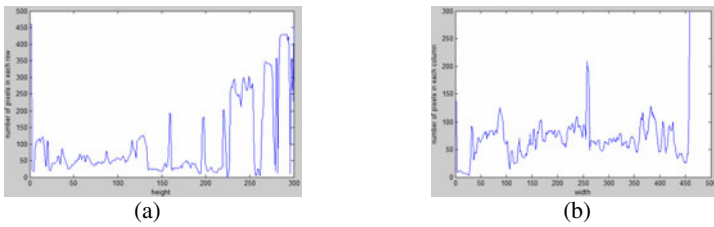


Fig. 5. (a) Horizontal Projection profile (b) Vertical Projection profile for image in Figure 4

Step-6. Obtain weak refined image in 0 and 90 orientations using morphological structure element and create the final refined image using subtraction of the refined image and weak refined image.

Step-7. In this step, with the help of connected component labeling operator, the long edge of final refined image is removed and 4-neighbor connected component are utilized. Finally every edge is uniquely labeled as a single connected component with its own defined component number.

Step-8. In this step, segment out non-text regions using major to minor axis ratio with help of Heuristic Filtering. Only those regions in the retained image which have an area greater than or equal to $1/20$ of the maximum area region and remove those regions which have Width/Height < 0.1 ratio. Retained image shown in figure 6.

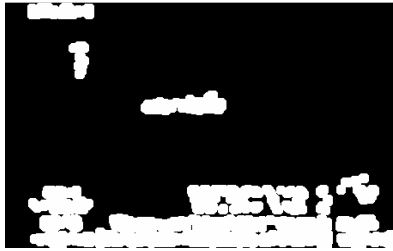


Fig. 6. Retained image

2.3 Text Extraction

The purpose of this section is to extract accurate binary characters from the localization text region.

Step-9. In this step a gap image will be originate by refine the localization of the region of detected text and then gap filling process is done. This shown in figure 7.



Fig. 7. Gap filling image

Step-10. Text segmentation is the next step to take place. It starts with extraction of text image from the gray image. Then, the segmentation process concludes with a procedure which enhances text to background contrast on the text image.

Step-11. The available common OCR system requires to easily recognized the character of an input images. Thus this process provides an output image with white text against a black background. Final text image shown in figure 8.

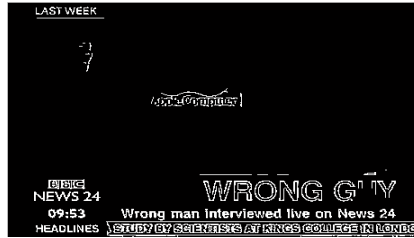


Fig. 8. Final text image

3 Result and Analysis

In order to evaluate the performance of the proposed method, there are 28 distinct test images are use which are of distinct font sizes, distinct perspective and distinct alignment under distinct circumstances. The results which are shown in figure 9 ~ 13 shows that our proposed method can detect the text with distinct font sizes, perspective, alignment, and detect the text string characters under distinct circumstances. The importance of algorithms testing with change of scale, lighting and orientation, is use to find the strength of every technique with change in these circumstance, and also use to find that where each technique is successful and where it fails.

Figure 9~13, show that our proposed method has excellent performance with wide variety of set of images. So that we can say our proposed method is a strong and impressive approach to find the text based images. The performance of each method has been calculated and it is based on obtained Recall rates and average time.

$$\text{Recall Reate} = (\text{Correctly Detected Words}) / (\text{Correctly Detected Words} + \text{False Negatives}) * 100 \quad (1)$$

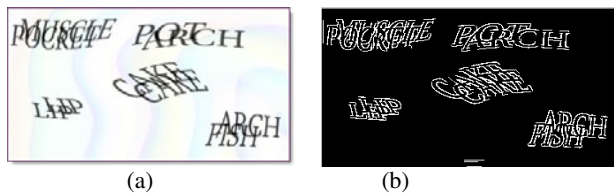


Fig. 9. Captcha image (a) Original image (b) Extracted image

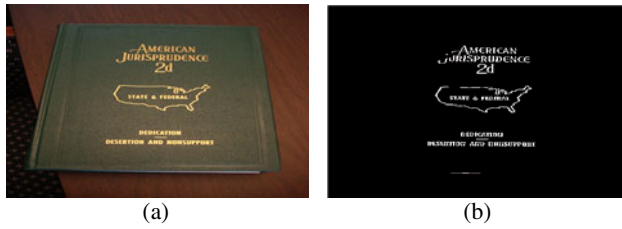


Fig. 10. Indoor Image (a) Original image (b) Extracted image



Fig. 11. Outdoor Image (a) Original image (b) Extracted image

The test set for this evaluation experiment consists of 28 single images selected randomly from the internet (Google search engine). The experiment is carried out on Matlab 7.0 software platform. The PC for experiment is equipped with an Intel P4 2.4GHz Personal laptop and 2GB memory. The total processing time, including read-in and write-out for all 28 images is less than 4 seconds.

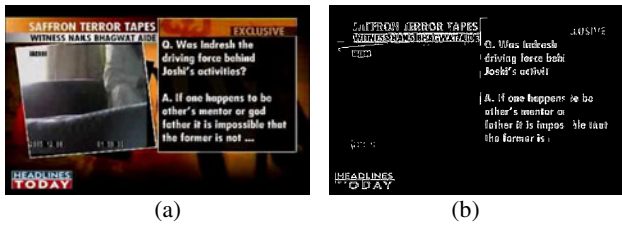


Fig. 12. News video frame (a) Original image (b) Extracted image



Fig. 13. Video frame (a) Original image (b) Extracted image

Table 1. Performance Comparison

Approach	Recall rate %	Average Time (s)
Proposed Approach	95.3	3.64
Xiaoqing et al. [2]	86.2	14.18
Gllavata et al. [3]	88.4	16.25
Li et al. [5]	91.1	12.9
Ye et al. [6]	90.8	10.1
Liu et al. [7]	91.3	11.7

Table 1 shows the performance comparison of our proposed method with several existing method, where our proposed method has a better performance in average time and recall rate. The reason for fast speed that the proposed method used line based edge approach cost less time.

4 Conclusions

In this paper, a fast and robust text extraction in images approach is proposed. The line detection edge based method in two directions is able to better represent the intrinsic characteristics of text. Experiment results show that our method can obtain 95.3 % recall rate and average text detection time is 3.64 second, which is superior to the existing text detection methods without much increasing the computational cost.

References

1. Al-Eidan, R.B., Al-Braheem, L., El-Zaart, A.: Line Detection Based On The Basic Masks And Image Rotation. In: 2nd International Conference on Computer Engineering and Technology, IEEE, pp. 465–469. IEEE, Chengdu (2010)
2. Liu, X., Samarabandu, J.: Multiscale Edge-Based Text Extraction from Complex Images. In: International Conference on Multimedia and Expo., ICME 2006, pp. 1721–1724. IEEE, Los Alamitos (2006)
3. Gllavata, J., Ewerth, R., Freisleben, B.: A robust algorithm for text detection in images. In: Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis, ISPA, vol. 2, pp. 611–616. IEEE, Los Alamitos (September 2003)
4. Liu, X., Samarabandu, J.: An edge-based text region extraction algorithm for indoor mobile robot navigation. In: International Conference on Mechatronics and Automation, pp. 701–706. IEEE, Los Alamitos (2005)
5. Li, X., Wang, W., Jiang, S., Huang, Q., Gao, W.: Fast and effective text detection. In: 15th IEEE International Conference on Image Processing, pp. 969–972 (October 2008)
6. Ye, Q., Huang, Q., Gao, W., Zhao, D.: Fast and robust text detection in images and video frames. *Image and Vision, Computing* 23, 565–576 (2005)
7. Liu, Q., Jung, C., Kim, S., Moon, Y., Kim, J.: Stroke filter for text localization in video images. In: Proc. Int. Conf. Image Process, Atalanta, GA, USA, pp. 1473–1476 (October 2006)

A Chip-Based Watermarking Framework for Color Image Authentication for Secured Communication

Soumik Das¹, Pradosh Banerjee¹, Monalisa Banerjee¹, and Atal Chaudhuri²

¹ MCA Department, Techno India, Salt Lake, India

² Computer Sc. & Engg. Department, Jadavpur University, India
soumikdas.techno@yahoo.co.in

Abstract. Multimedia contents have received a sharp attention now a day as a result of massive development of cyber crime. Illegal replication, violation of authentication, misappropriation of digital content has achieved an enigmatic growth. In this respect we've already proposed different schemes for invisible watermarking for color image authentication in fragile domain. In current communication we are going to propose a chip-based framework for color image authentication in accordance with our prescribed LSB scheme. In this watermarking scheme, a secret key will be given to the user and with the help of a particular hash function the user can authenticate the ownership of an image. With the aforesaid secret key and hash function the secured communication is guaranteed. The original watermark will not be retrieved if any intruder tries to perform watermark extraction with an inappropriate key and a incorrect hash function. We named our proposed chip as BLIND CHIP, because we've used blind method for watermark extraction where neither the host image nor the watermark image is needed at the time of watermark extraction.

Keywords: access controls, authentication, cryptographic controls, and verification.

1 Introduction

Digital Image Watermarking is the technique for embedding information into a digital image. The image watermarking framework provides a persistent connection between the authenticator and the image it authenticates [1]. Our focused area is Invisible Fragile Watermarking. In this kind of watermarking the information is added as digital data to the original, but it will not be visible in Human Visual System (HVS). Our already proposed different frameworks have been proven to embed color watermark to color host images in a very efficient manner [3] [4]. In the current communication we want to give completeness and total effectiveness to our proposed LSB framework. We are proposing a digital IC based hardware solution comprising of multiplexers, decoders, counters, shift-registers etc. separately for embedding and extracting the watermark. Through our BLIND CHIP a user can authenticate image without performing a program compilation. The speedup ratio of hardware platform over the software platform will definitely be increased. With addition to that security issue will be proven efficient using a chip rather than a program. In current communication we've discussed the framework and algorithmic approach for embedding and extracting in the next section.

2 Algorithm and Chip-Based Approach

Our already proposed framework will extract LSB's of 'Blue' value from each pixel of individual blocks of host image using a secret key and a hash function. 'Blue' is the most high frequency part of an image. That's why it is less sensitive to Human Visual System [2]. We've only changed the 'Blue' value of the color host image. The 'Red' value and 'Green' value of color host image are remaining unchanged.

Let the color host image be 'Y'. A color watermark 'A' will be inserted in it and again will be extracted from it for authentication. The color host image is divided into several no. of equal blocks (Y_i) according to size of the image, where each block size is same with color watermark 'A'. As per our LSB scheme we need such twenty-four blocks. In the following discussion, we concentrate on inserting and extracting all bit information of color watermark 'A' into the corresponding color host image block Y_i and form watermark embedded image block W_i .

[insert_color_watermark]

Input: Color host image block (Y_i), Watermark image (A), Secret key (K).

Output: Watermark embedded image block (W_i)

Step1: $Y_{ip} = \text{get_pixel_info}(Y_i)$.

Step2: $L_i = \text{onebit_to_zero}(Y_{ip})$.

Step3: $H_i = \text{hash_to_generate_mesg_digest}(L_i, K)$.

Step4: $A_i = \text{extract_onebit}(A)$.

Step5: $X_i = \text{compute_XOR}(H_i, A_i)$

Step6: $W_i = \text{replace}(Y_{ip}, X_i)$.

Programmatic functions are described hereinafter.

get_pixel_info (Y_i): This function will bring the binary information of blue value of a pixel of host image block (Y_i) that forms Y_{ip} .

onebit_to_zero (Y_{ip}): This function will convert the LSB of the binary values of Y_{ip} to zero to form L_i .

hash_to_generate_mesg_digest (L_i, K): This function will generate the message digest (H_i) depending on the value of L_i and a 16-bit secret key K.

extract_onebit (A): This function will extract LSB of each pixel of watermark image (A) that forms A_i .

compute_XOR (H_i, A_i): This function will perform XOR operation between H_i and A_i to form X_i .

replace (Y_{ip}, X_i): This function will replace the LSB of Y_{ip} with the value X_i which we got by XORing H_i and A_i to form W_i .

As per chip based solution for embedding watermark into the host we are considering a block (Y_i) of size 2X2 of image Sample1.jpg as our host image and Logo.jpg as our watermark image (A). A secret key (k) is also defined. Now we are stepping forward according to our proposed chip-based solution. We've used four IC74151 (8:1 MUX) to multiplex the eight individual bits of 'Blue' value of the host image for a 2X2 block. The ICs are denoted as M3, M4, M5, and M6. The output lines of IC74151 (3-to-8 line Decoder) are used to trigger the multiplexers one by one. A 2-bit counter (C1) enabled by a clock will provide the input of the IC74151 (D1). By providing 00

at the selection lines of all the four IC74151, we can multiplex the LSB of the ‘Blue’ bit pattern for all pixels. And then those least significant bits are stored in IC74195 (4-bit Shift Register R1).

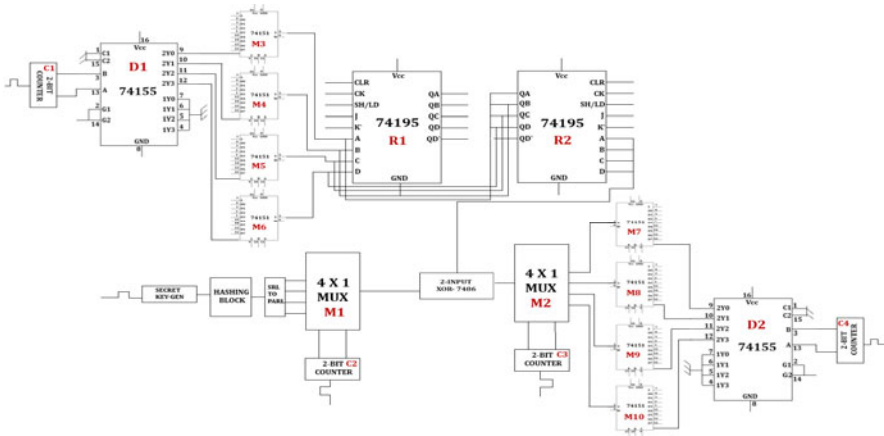


Fig. 1. BLIND CHIP for Embedding the Watermark

Hash output will be generated with the help of a secret key generator and a serial to parallel converter. The parallel outputs of converter are applied as inputs of another 4X1line multiplexer (M1) enabled by a clocked 2-bit counter (C2). Now with every clock pulse the multiplexers M1 and M2 multiplexes their input lines at their output one by one. We’ve used an IC7486 (2-input XOR). The output lines of M1 and M2 are applied as input lines of IC7486 for bitwise XORing. The resultant information is stored in another IC74195 (R2). After that the content of R1 will be replaced by content of R2 through intermediate communication lines. Fig.1 shows the chip-based solution for embedding watermark.

To extract the watermark from watermarked image we’ve prescribed an algorithm given below:

[extract_color_watermark]

Input: Watermark embedded image block (W_i), Secret key (K).

Output: Extracted watermark image block (A'_i).

Step1: $L'_i = \text{onebit_to_zero}(W_i)$.

Step2: $H_i = \text{hash_to_generate_mesg_digest}(W_i, K)$.

Step3: $R_i = \text{retrieve}(W_i)$.

Step4: $A'_i = \text{compute_XOR}(H_i, R_i)$.

retrieve(W_i): This function will retrieve LSB of blue value of each pixel of watermarked image (W_i) that forms R_i .

Other functions are already defined during embedding.

As per chip based solution for extracting watermark from the host image we are considering a 2X2 block of the watermarked image (W_i). We’ve used four IC74151 (8:1 MUX) to multiplex the eight individual bits of ‘Blue’ value of the watermarked

image. The ICs are denoted as M3, M4, M5, and M6. The output lines of IC74155 (3-to-8 line Decoder) are used to trigger the multiplexers one after another. A 2-bit counter (C1) enabled by a clock will provide the input of the IC74155 (D1). The output lines of the multiplexers are employed as input lines of a 4X1 line multiplexer (M2) which is triggered through a clocked 2-bit counter (C2). By providing 00 at the selection lines of all the four IC74151, we can multiplex the LSB of the ‘Blue’ bit pattern for all pixels of the watermark image.

On the other hand the hash output will be generated with the same secret key, a serial to parallel converter and a 4X1line multiplexer (M1). The out put lines of M1 and M2 will be employed as input lines of IC7486 (2-input XOR). The output line of the 2-input XOR IC will be applied to a IC74195 (4-bit Shift Register R1). After bitwise XORing the individual bits will be stored in R1. Fig.2 shows the chip-based solution for extracting watermark.

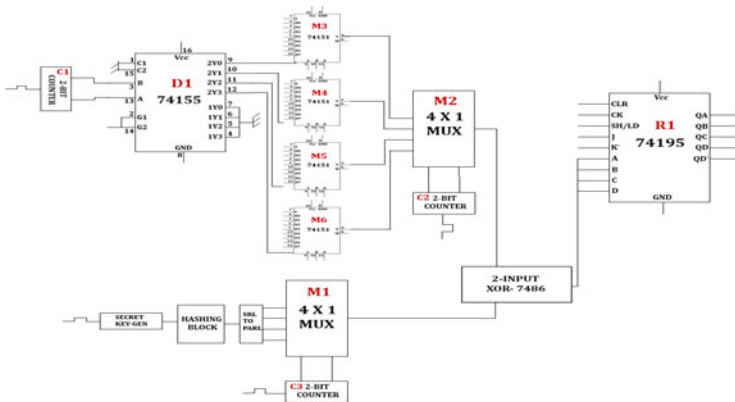


Fig. 2. BLIND CHIP for Extracting the Watermark

3 Algorithm Execution

In this section we are going to show, how the prescribed technique is applied on a color host image for embedding watermark and extracting watermark using “insert_color_watermark” and “extract_color_watermark” those are defined in the previous section. We are taking a 2X2 block (Y_i) of our host image Sample1.jpg and another 2X2 block (A) of our watermark image Logo.jpg. Now we are stepping forward with the algorithm execution for embedding watermark. The RGB information of 2X2 block of our host image Sample1.jpg is given below by Y_i . The corresponding ‘Blue’ information of aforesaid 2X2 block represented by Y_{ip} is applied as input lines of M3, M4, M5 and M6. By providing 00 at the selection lines of the multiplexers, we’ll be able to store the LSB information of the ‘Blue’ bit streams into R1. The Y_{ip} is given below.

$$\begin{aligned}
 Y_i &= \begin{bmatrix} 155\&153\&154 & 153\&153\&153 \\ 152\&159\&153 & 150\&152\&156 \end{bmatrix} & Y_\Psi &= \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & & & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & & & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{bmatrix} \\
 L_i &= \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & & & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & & & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{bmatrix} & A &= \begin{bmatrix} 172\&197\&168 & 154\&183\&164 \\ 155\&186\&156 & 142\&173\&141 \end{bmatrix} \\
 A_r &= \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & & & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & & & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} & A_i &= \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}
 \end{aligned}$$

On the other hand secret key generator will produce a 16-bit key value, K=1001 0011 0101 0011. Depending on that we are having a hash output H_i is given below.

$$\begin{aligned}
 H_i &= \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} & X_i &= \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} & W_i &= \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & & & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & & & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix}
 \end{aligned}$$

At the extraction end we are taking W_i as our input. The bit streams are employed as input lines of M3, M4, M5 and M6 which multiplex the LSB information of bit streams. The output lines of those multiplexers are connected as input lines of M2. After four clock pulses we are having R_i which is applied to the one input line of IC7486. Now we are stepping forward with the prescribed algorithm.

$$\begin{aligned}
 W_i &= \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & & & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & & & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix} & R_i &= \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \\
 L'_i &= \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & & & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & & & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{bmatrix}
 \end{aligned}$$

On the other hand secret key generator will produce the same 16-bit key value, K=1001 0011 0101 0011. That's why we are having the same hash output H_i is given below.

$$\begin{aligned}
 H_i &= \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} & A'_i &= \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}
 \end{aligned}$$

In the next section the result set analysis of our proposed framework is reported.

4 Experimental Result Set Analysis

Our proposed scheme is carried out with detailed experiment. We've tested proposed framework on several different color images and we've found a set of excellent result set. Table 1 shows one of the result sets and some other tested results are depicted in Table 2. From Table 1 it is clear that after embedding the color watermark to our color host image, the watermark is entirely invisible to the HVS. And 3rd column of Table.2 for LSB scheme is provided in this regard. In this regard we've used a measure called PSNR [1] [3]. The PSNR is most commonly used as a measure of quality of

watermarked image. It is most easily defined via the root mean squared error (RMSE) [1]. We've also obtained the highest value for Normalized Correlation (nc) [2], i.e., unity which conforms our extracted watermark retained as it is. Tested nc values are depicted in 4th column of Table 2.

Table 1.

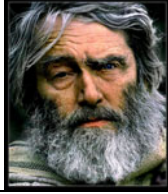



Host Image	Watermark
	
Watermarked Image	Extracted Watermark
	

Table 2.

Image Name	Size	PSNR	nc
apple.jpg	256 * 256	81.04	1
animal.jpg	500 * 375	84.13	1
river.jpg	615 * 413	86.01	1
jesus.jpg	625 * 415	87.26	1
nature.jpg	1024* 768	89.03	1
old_man.jpg	1000 * 1500	90.94	1
sky_blue.jpg	1400 * 1350	90.96	1

5 Conclusion

The detailed experiment has been performed and it is found that our LSB framework is able to embed color watermark to color host images in an efficient manner and perceptually the watermark remains invisible in the watermarked image. Through the chip based solution the ownership authentication is guaranteed in an efficient way and reaches us to greater horizon such as mobile communication. Having chip based solution the portability is also enhanced. We've also ensured that the extracted watermark remains intact. And we've achieved such excellent PSNR values.

References

1. Cox, J., Miller, M.L., Bloom, J.A., Fridrich, J., Kalker, T.: Digital Watermarking and Steganography, 2nd edn. Morgan Kaufmann Publishers, San Francisco (2008) ISBN:978-0123725851
2. Parek, R.: Principle of Multimedia. Tata McGraw Hill, New Delhi (2006) ISBN: 0-07-058833-3
3. Das, S., Banerjee, P., Paul, S., Sinha Ray, A., Banerjee, M.: Pioneering the Technique for Invisible Image Watermarking on Color Image. International Journal of Recent Trends in Engineering, 508–511 (May 2009) ISSN: 1797-9617
4. Das, S., Banerjee, P., Paul, S., Sinha Ray, A., Banerjee, M.: A New Introduction towards Invisible Image Watermarking on Color Image. In: Published in Proceedings of IEEE International Advance Computing Conference (March 2009) ISBN:978-1-424429271

An Image Decomposition Parallel Approach to Train Flann for an Adaptive Filter

Manoj Kumar Mishra¹, Nachiketa Tarasia¹,
Bivasa Ranjan Parida², and Sarita Das³

¹ KIIT University

Bhubaneswar, Orissa, India

manojku.mishra05@gmail.com, nachiketa.tarasia@yahoo.com

² S.T.A.R., Taraboi, Jatni, Khurda

bivasranjanparida@gmail.com

³ EAST, Phulnakhara, Bhubaneswar

das_sarita4u@yahoo.com

Abstract. Filtering method is applied to the images corrupted at the time of transmission due to several noises, with varying strengths and different noise probability. Neural network based image filter is one of the most important example of adaptive image filter. Adaptive neural network filter remove various types of noise such as Gaussian noise and impulsive noise. Neural networks are based on the concept of training or learning by examples and have already been applied in several domains of image processing including image filtering. But training of those neural networks consume much time before it is actually tested on such as image filtering. Applying parallelism to image processing is increasingly practical and necessary, as our desktops are becoming multicore machines replacing single core. Therefore, this paper proposes a parallel approach named image decomposition parallel approach to train FLANN (Functional Link Artificial Neural Network). Well trained FLANN is used for rectifying the corrupted pixels to restore the image. Experimental results obtained through SPMD(Single Program Multiple Data) simulation environment show that the proposed parallel approach to train the FLANN is feasible as it substantially reduces the training period and also make it an efficient filter to restore the image fairly well maintaining the quality of the filtered image. Hence, this method is suitable for real time image restoration applications.

Keywords: FLANN, salt and pepper, SPMD, Image decomposition.

1 Introduction

Artificial Neural Network techniques have gained popularity among many researchers and its application in image processing is one of the promising research fields. Neural network based image filtering technique is applied to the image that gets noisy at the time of transmission and percentage of noise varies from time to time and also changes in a fraction of second. So, it is difficult to choose

a filter at that small time window. To avoid different limitations of fixed filters, adaptive filters are designed that adapt themselves to the changing conditions of noise. In such an application the image filter must adapt the image local statistics, the noise type and it must adjust itself to change its characteristics so that the overall filtering improves substantially.

Application of Artificial Neural Networks has been reported in several domains of image processing including image filtering. The MLP is a multilayer architecture with one or more hidden layer(s) between its input and output layers. All the nodes of a lower layer are connected with all the nodes of the adjacent layer through a set of weights. All the nodes in all layers (except the input layer) of the MLP contain a nonlinear $\tanh()$ function. A pattern is applied to the input layer, but no computations takes place in this layer. Thus, the output of the nodes of this layer is the input pattern itself. The weighted sum of outputs of a lower layer is passed through the nonlinear function of a node in the upper layer to produce its output. Thus, the outputs of all the nodes of the network are computed. The outputs of the final layer (output layer) are compared with a target pattern associated with the input pattern. The error between the target pattern and the output layer node is used to update the weights of the network. The mean square error (MSE) is used as a cost function. Due to the multilayer architecture, the MLPs are inherently computationally intensive. Although the MLP provides robust solution, its excessive training time and high computational complexity appear as two major drawbacks of this approach. The functional link artificial neural network (FLANN) by Pao [1] can be used for function approximation and pattern classification with faster convergence and lesser computational complexity than a MLP network. A FLANN using $\sin()$ and $\cos()$ function for functional expansion for the problem of nonlinear dynamic system identification has been reported [2]. But practically, large training time of FLANN makes it unattractive for implementation. However, further research is desirable for reducing the training period substantially and to devise an efficient filter for effective suppression of noises and fairly restoring the image.

Image restoration methods generally model the degradation process and apply an approximately inverse process to the degraded image to recover the original image [3]. The effectiveness of such image restoration techniques depends on the availability and completeness of knowledge about the degradation process as well as on the structure of the processing scheme implementing the restoration. Various image restoration methods have been proposed in the literature on digital image processing. Descriptions of image restoration techniques can be found in books on image processing such as [4]. Traditional image restoration methods are based on linear processing of image signals. Weiner filters [5] and recursive (Kalman) filters [6] fall under this category. Applying parallelism in solving typical problems having huge computation is becoming trend in almost all research fields. Parallelism can be applied in training of NN, which has not been reported in any of the previous research works in the field of image filtering by decomposing the image into p number of partitions where p is some power of 2. In this paper FLANN is trained on each input part of the decomposed image matrix

independently in the SPMD simulation environment. The objective of this paper is to reduce training time, error rate and enhance the quality of the filtered image.

The paper is further organized as follows. In section 2 contains structure of FLANN and section 3 describes the learning principle of FLANN. Section 4 contains the description of proposed parallel approach for training of FLANN. Section 5 provides simulation and experiments of the proposed approach with help of Matlab 7.9 toolkit and its results are compared to serial approach. Some final conclusions are drawn in section 6 and lastly references.

2 Structure and Learning of FLANN

Recently artificial neural network (ANN) has emerged as a powerful learning technique to perform complex tasks in highly nonlinear environment [7]. The FLANN, which is initially proposed by Pao [1], is a single layer ANN structure capable of performing complex decision regions by generating nonlinear decision boundaries. The structure of an FLANN is shown in Fig. 2. FLANN is a single-layer flat structure where the need of hidden layers is removed by transforming the input patterns to a higher dimensional space as a result the patterns in the projected higher dimensional space become linearly separable. FLANN performs pattern enhancement by using a set of orthogonal functions of either an element or the entire pattern and is capable of representing nonlinear mapping between the input and output spaces [7]. Usually, in an FLANN, the functional expansion is carried out by using trigonometric functions [8,9,10].

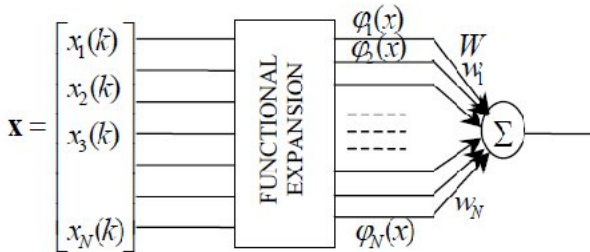


Fig. 1. Structure of FLANN with a single output

The learning process involves updating of the weights of FLANN in order to minimize a given cost function. An artificial neural network can approximate a continuous multi variable function $f(x)$. Let the approximating function be represented as $f_w(x)$. In the FLANN, a set of basis function ϕ and a fixed number of weight parameters W are used to represent the $f_w(x)$. With a suitable set of basis functions, the problem is then to find the weight parameters W that provides the best possible approximation of f on the the set of input-output examples. The well known back propagation algorithm is used to update the

weights of FLANN. The use of FLANN for the purpose of multidimensional function approximation has been discussed in [11].

Referring to fig.1, let there be K number of input-output pattern pairs to be learned by the FLANN. Let the input pattern X is of dimension and for ease of understanding, let the output, y be a scalar. Each of the input patterns is passed through a functional expansion block producinh a corresponding N -dimensional ($N \geq n$) expanded vector.

In this case, the dimension of the weight matrix is of $1 \times N$ and hence, the individual weights are represented by a single subscript. Let $w = [w_1 w_2 \dots w_N]$ be the weight vector of its FLANN. The linear weighted sum, S_k is passed through the $\tanh(\cdot)$ non-linear function to produce the output $\hat{y}(k)$ with the following relationship:

$$\hat{y}k = \tanh(S_k), \text{ or, } S_K = \frac{1}{2} \log_e \left(\frac{1 + \hat{y}k}{1 - \hat{y}k} \right) \tag{1}$$

Let K number of patterns be applied to the network in a sequence repeatedly. Let the training sequence be denoted by X_k, Y_k and the weight of the network be $W(k)$, where k is the discrete time index given by $k=k+\lambda k$, for $\lambda=0,1,2,\dots$, and $k=0,1,2,\dots,K$. At k^{th} instant, the n -dimensional input pattern and the m -dimensional FLANN output are given by $X_k = [x_1(k)x_2(k)\dots x_n(k)]^T$ and $\check{y}(k) = [\check{y}_1(k)\check{y}_2(k)\dots\check{y}_m(k)]^T$ respectively. It's corresponding target pattern is represented by $Y(k) = [y_1(k)y_2(k)\dots y_m(k)]^T$. The dimension of input pattern increases from n to N by a basis function ϕ given by $\phi(X_k) = [\phi_1(X_k)\phi_2(X_k)\dots\phi_N(X_k)]^T$. The $(m \times N)$ dimensional weight matrix is given by $W(k) = [W_1(k)W_2(k)\dots W_m(k)]^T$ where $W_j(k)$ is the weight vector associated with j^{th} output and is given by

$$W_j(k) = [W_{j1}(k)W_{j2}(k)\dots W_{jN}(k)] \tag{2}$$

The j^{th} output of the FLANN is given by

$$\hat{y}_j(k) = \rho \left(\sum_{i=1}^N w_{ji}(k)\phi_i(X_k) \right) = \rho (w_j(k)\phi^T(X_k)) \tag{3}$$

for $j=1, 2, 3,\dots,m$.

Let the corresponding error be denoted by $e_j(k) = y_j(k) - \check{y}_j(k)$. Using the back propagation algorithm for single layer, the update rule for all the weights of the FLANN is given by

$$W(k + 1) = W(k) + \mu\delta(k)\phi(X_k). \tag{4}$$

where $W(k) = [W_1(k)W_2(k)\dots W_m(k)]^T$ is the $m \times N$ dimensional weight matrix of the FLANN at the k^{th} time instant, $\delta(k) = [\delta_1(k)\delta_2(k)\dots\delta_m(k)]^T$, and $[\delta_j(k) = (1 - \bar{y}_j(k)^2)e_j(k)]$.

3 Proposed Parallel Approach

The values of a digital image are represented in a two-dimensional Matrix form and any changes in values of the matrix cause noise in the image. The proposed parallel training approach of FLANN as shown in the fig.3 is done by one-dimensional image decomposition technique, where FLANN training is performed by distributing decomposed blocks of the image matrix among multiple CPUs available in multiprocessors system for parallel execution. After decomposition of some reference image, here the FLANN is trained on each input block of the image matrix in parallel and then updated weights are used to test on individual part of the image in parallel and are tested on some other images. This training is performed for required number of iterations thus producing error convergence set for each input block. The corresponding updated weights generated by FLANN whose error values converges to zero is taken to test the entire image. It is found that the proposed NN can be trained on any given image parallelly thus reducing the training time and the best weights obtained after training can be used cancel noise adaptively from all other noisy images.

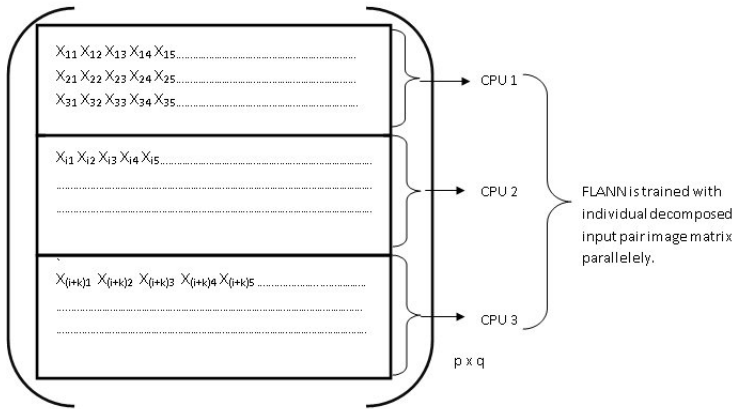


Fig. 2. An example of one-dimensional image decomposition

Three basic computations, i.e., addition, multiplication and computation of $\tanh(\cdot)$ are involved for updating weights of the ANN. The computations in the network are due to the following requirements:

1. Forward calculations to find the activation value of all the nodes of the entire network;
2. Back-error propagation for the calculation of square error derivatives and
3. Updating the weights of entire network.

Total number of weights to be updated in one iteration of FLANN is $(n_0 + 1)n_L$ Where n_0 and n_L are the number of nodes in the input and output layer, respectively. It can be noted that there is no hidden layer in FLANN. In case of

proposed parallel approach the total number of weights updation is $(n_0 + 1)pn_L$, where p is the number of image partitions assuming p number of processor are available for each partition. But total number of weights updations in all iteration remains same. Therefore, it reduces the training time for FLANN. The primary purpose of this paper is to reduce training time, error rate and better quality of the filtered image with respect to corrupted image.

4 Simulation and Experimental Studies

The simulation environment is created with the help of MatLab ver7.9, Pentium dual core processor, 1.6 GHz, and 512MB RAM. The experimental studies has been conducted in SPMD(Single Processor Multiple Data) parallel platform of MatLab. Here the images used for simulations are text.tif and rice.tif shown in fig. 3 and 4 respectively, which are standard grayscale images of size 256x256. Experiments were performed to evaluate the performance of the filter by decomposing the image matrix parallelly using Functional link neural network with Trigonometric functional expansion. In this simulation, the image matrix corrupted by salt and pepper noise of density 0.05 is decomposed into four independent parts. Therefore, FLANN can be trained on each individual decomposed image matrix parallelly as shown in fig.2.

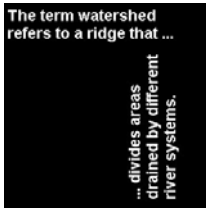


Fig. 3. Text.tif

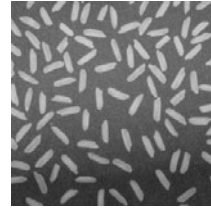


Fig. 4. Rice.tif

Here the 9 inputs of this network are the 3 x 3 window of the noisy image and the target was the middle value of the matrix window. Figure 5 represents an image in matrix format. So here the first 3 x 3 window is selected as input, and target is selected as the middle value of the window, i.e. x_{22} . Here, this FLANN training process is carried out independently on each individual decomposed image matrix.

We process the window iteratively in order to cover the image matrix. These 9 pixels values are given as input to the FLANN and it is then expanded up to 45 patterns using trigonometric expansion. The initial weights have been taken from the range of -0.5 to +0.5 and randomly distributed between 45 layers. The network is trained by various parameters like bias=1.0, learning rate=0.001.

In the first experiment FLANN is trained on the total corrupted 'text.tif' serially without decomposition and in the next experiment it is trained by the

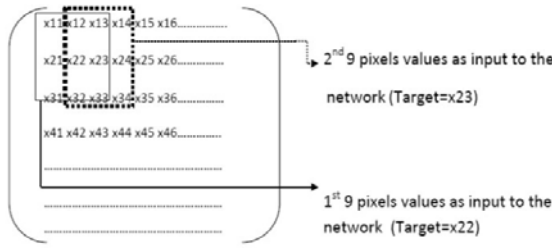


Fig. 5. One-dimensional image decomposition

proposed parallel decomposition method described above. After getting the desired weight set, it is tested on another image ‘rice.tif’ corrupted with salt & pepper of density 0.05. After getting the filtered image, the NRDB(Noise Reduction in DB) is being calculated for both the methods. MSE(Mean Square Error) and NRDB are given by

$$MSE = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} ||I(i, j) - K(i, j)||^2 \tag{5}$$

$$NRDB = 10 * \log_{10}(MSE_{in}/MSE_{out}). \tag{6}$$

MSE_{in} :- Error between original and noisy image. MSE_{out} :- Error between original and filtered image.

The simulation results of both experiment is given below in table 1. It can be seen that the time taken to train FLANN in proposed parallel method is nearly 50% less by distributing the image matrix equally among 4 different processes as compared to serially trained FLANN. It can also be observed that parallelly trained FLANN shows lower MSE and higher NRDB as compared to the other.

After decomposing the image into four blocks, each individual block of image matrix is given to four different CPUs, where the FLANN algorithm is trained parallelly. FLANN algorithm running in each processor updates its weight parameter independently by calculating the error between the training and the target data set. Here each FLANN is trained for 60 iterations. It can be observed from the above graph, CPU2 error converges to zero and remaining constant thereafter. The weights generated at CPU2 are taken for testing the ‘rice.tif’ corrupted image as shown in fig.7 and the following filtered image is obtained as shown in fig.8.

Speed-up: $S_P = T_S/T_P = 861.48511/495.86157=1.7373$, T_S is the serial execution time, T_P is the parallel execution time and P is the number of processes.

Efficiency: $E_P = S_P/P=1.7373/4= 0.4343$, where S_P is the speed-up, P is the number of processes.

Table 1. Simulation results

	Serial method	Parallel method
Time taken(sec)	861.48511	495.86157
MSE _{in}	0.0145	0.0145
MSE _{out}	0.0112	0.0065
NRDB	0.4376	2.1742

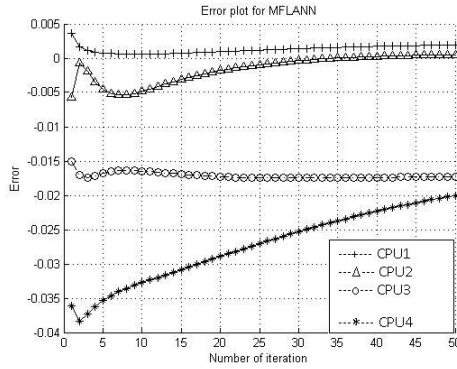


Fig. 6. Convergence characteristics of FLANN at four different CPUs

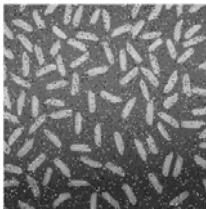


Fig. 7. Noisy image

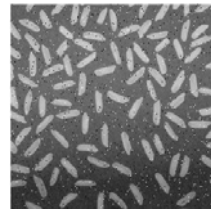


Fig. 8. Filtered image

5 Conclusion

The proposed approach of training and then filtering an image with FLANN is effective as it attains a speed-up of almost 1.7, and efficiency of 0.4343. This filter is also capable of restoring the image to an acceptable level by suppressing the noise. Because of its lower execution time requirement and ability to perform complex mapping between multi-dimensional input and output spaces, this parallel approach has the potential applications in online image processing and may be used in other areas of science and engineering. The proposed parallel decomposition approach outperforms the serial one, both qualitatively and

quantitatively. Further work is necessary to address other issues such as scalability, load balance efficiency and generalizing this filter applicable to other types of noise.

References

- [1] Pao, Y.-H.: Adaptive pattern recognition and neural networks. Addison-Wesley, Reading (1989)
- [2] Patra, J.C., Paul, R.N., Chatterji, B.N., Panda, G.: Identification of nonlinear dynamic systems using functional link artificial neural networks. *IEEE Trans., Systems, Man and Cybernetics, Part B* 29, 254–262 (1999)
- [3] Cha, I., Kassam, S.A.: RBFN restoration of nonlinearly degraded images. *IEEE Trans. on Image Processing*, 5, 964–975 (1996)
- [4] Jain, A.K.: Fundamentals of Digital Image Processing. PHI, Englewood Cliffs (1990)
- [5] Chellapa, R., Kashyap, R.L.: Digital image restoration using spatial interactive models. *IEEE Trans. on Acoust., Speech, Signal Processing ASSP-30*, 361–472 (1982)
- [6] Dikshit, S.S.: Recursive Kalman window approach to image restoration. *IEEE Trans. on Acoust., Speech, Signal Processing ASSP-30*, 125–129 (1982)
- [7] Haykin, S.: Neural networks. Maxwell Macmillan, Ottawa (1994)
- [8] Patra, J.C., Pal, R.N.: A functional link artificial neural network adaptive channel equalization. *Signal Process.* 43(2), 181–195 (1995)
- [9] Patra, J.C., Pal, R.N., Baliarsingh, R., Panda, G.: Nonlinear channel equalization for QAM signal constellation using artificial neural networks. *IEEE Trans., Systems, Man and Cybernetics, Part B* 29(2), 262–271 (1999)
- [10] Weng, W.D., Yang, C.S., Lin, R.C.: A channel equalizer using reduced decision feedback Chebyshev functional link artificial neural networks. *Inf. Sci.* 177, 2642–2654 (2007)
- [11] Patra, J.C., Pal, R.N., Chatterji, B.N., Panda, G.: Identification of nonlinear dynamic system using functional link artificial neural networks. *IEEE Trans., Systems, Man and Cybernetics, Part B* 29(2), 254–262 (1999)

Palette Based Technique for Image Steganography

Anuradha Lamgunde¹ and Achana Kale²

¹ Information Technology Dept., Don Bosco Institute of Technology,
Kurla (West), Mumbai, India

² Information Technology Dept., Thadomal Shahani College of Engg.,
Bandra (West), Mumbai, India
anuradha01.02@gmail.com, archanatsec@yahoo.co.in

Abstract. Today worldwide, security and data hiding concerns are increasing day by day. Various security devices, systems and algorithms are used at places for the same. Most of these systems are implemented at places like airports, military areas, private or government offices.

Steganography means covered or hidden writing. The objective of steganography is to send message through some innocuous carrier. The message to be sent could be a text, an image or an audio file. Steganography techniques prevent the fact that a secret message is being sent at all.

Steganographic security is mostly influenced by the type of cover media; the method for selection of places within the cover that might be modified; the type of embedding operation; and the number of embedding changes that is a quantity closely related to the length of the embedded data. Given two embedding schemes that share the first three attributes, the scheme that introduces fewer embedding changes will be less detectable.

Keywords: Steganography, Palette, Data Hiding.

1 Introduction

“Palette Based Technique for Image Steganography” is the scheme that uses image as the carrier medium to hide secret messages. Secret message is encrypted before hiding. The bits of secret message will be hidden inside the cover image using the stretched palette of the cover image.

The aim of the work is to implement Palette based technique for Image Steganography. For this a technique is considered that will work on images as well as text messages except TIFF images. Also for Text messages user has to enter the text i.e., it will not work for text documents. It is also able to encrypt the secret message using a strong encryption algorithm and hide secret messages inside the cover image using the stretched color palette of the cover image. Use of both steganography and cryptography techniques result in highly secure applications.

The process of stretching palette by adding new colors is an overhead & takes some time. However, the time is not a critical factor for this type of work.

2 Literature Review

Steganography has its roots in military and government applications and has advanced in genuity and complexity. Steganography is a two part of greek origin: "Stegano"-Covered and "graphy"-writing. Herodotus's histories describe two types of early steganography. The first type involved shaving of a slave's head, then a tattoo was inscribed on the scalp. When the slave hair had grown back and hidden the message the slave was send to warn of the Persian's impending invasion. For retrieval of information once again the slave's head was shaved by the recipient. Another method was to modify ancient writing tablets. The layer of wax covering the tablet was the surface upon which the messages were written. Steganography has taken a giant leap forward which started in the 1990's when the governments, industries, private citizens, and even the terrorist organizations began using steganography to embed messages and photos into various types of media(digital, audio files, etc.) Some of the techniques are:

LSB Method inserts the message bits in the carrier bit stream, substituting insignificant information in a carrier file with the secret data. The least significant bit of information at each image pixel is replaced with a bit from the hidden message. [2]

In Cover selection based on similarity of image blocks the blocks of secret image are compared with blocks of set of cover images and the image with the most similar block to those of the secret image is selected as a best candidate to carry secret image. [1]

A JPEG Compression Resistant Steganography scheme for Raster Graphic Images hides data in bitmap images, in a way that there is almost no perceptible difference between the original image and this new stego image and it is also resistant to JPEG compression. JPEG compression is performed independently on blocks of 8x8 pixels in an image while converting it to the JPEG format. The proposed scheme makes use of this property. [4]

3 Description

3.1 Indexed Color and Palettes

File formats like TIF and JPG store a 24 bit RGB value for each of the millions of image pixels. But GIF files only store a 4 or 8 bit index at each pixel, so that the image data is 1/6 or 1/3 the size of 24 bits.

Indexed Color is limited to 256 colors, which can be any 256 from the set of 16.7 million 24 bit colors. Each color used is a 24 bit RGB value. Each such image file contains its own color palette, which is a list of the selected 256 colors (or 16 colors in a smaller palette). Images are called indexed color because the actual image color data for each pixel is the index into this palette. Each pixel's data is a number that specifies one of the palette colors, like maybe "color number 82", where 82 is the

index into the palette, the 82nd color in the palette list of colors. The palette is stored in the file with the image.

The index is typically a 4 bit value (16 colors) or 8 bit value (256 colors) for each pixel, the idea being that this is much smaller than storing 24 bits for every pixel. But an 8 bit number can only contain a numerical value of 0 to 255, so only 256 colors can be in the palette of possible colors.

The file also contains the palette too, which is the table of the selected 24 bit colors, or 3 bytes of RGB overhead for each color in the palette (768 bytes for 256 colors). The first RGB color in the table is index 0, the second RGB color is index 1, etc. There can be at most only 256 colors in the palette.

So indexed files have 24 bits stored for each palette color, but not for each pixel. Each pixel only stores either a 4 bit or 8 bit index to specify which palette color is used. There are various ways to create the palette, to choose the possible color choices that it will contain.

Converting to 16 or 256 colors

There are several ways to convert to indexed color. Two choices are required, to specify a palette of colors, and also a choice how to dither or show colors not in that limited palette.

3.2 Characterizing Data Hiding Techniques

Steganographic techniques embed a message inside a cover; various features characterize the strengths and weakness of the methods.

Hiding capacity. Hiding capacity is the size of information that can be hidden relative to the size of the cover. A larger hiding capacity allows the use of smaller cover for a message of fixed size, and thus decrease the bandwidth required to transmit the stego-image.

Secrecy. A person should not be able to extract the covert data from the host medium without the knowledge of the proper secret key used in the extraction procedure.

Imperceptibility. The medium after being embedded with the covert data should be indiscernible from the original medium. One should not become suspicious of the existence of the covert data within the medium.

Accurate Extraction. The extraction of the covert data from the medium should be accurate and reliable.

Tamper Resistant. Beyond robustness to destruction, tamper resistant refers to the difficulty for an attacker to alter or forge a message once it has been embedded in a stego-image, such a pirate replacing a copyright usually also demand a strong degree of tamper resistant.

Other Characteristics. Computational complexity of encoding and decoding is another consideration and individual applications may have additional requirements.

3.3 Shortcomings of Steganography

Because Steganography has gained popularity only in the past decade, there are many flaws and vulnerabilities that still need to be addressed. Consequently, new Steganography technologies are being released with frequency.

- **Reveling the existence of hidden data**

Because steganography modifies an existing file that is most likely in circulation on the internet, a bitwise comparison of a given file with the “same” file suspected of containing hidden information can reveal use of steganography. Additionally, two communicating parties can be easily identified as communicating covertly if files that normally would not be exchanged suddenly are. For example, two business executives frequently exchanging photographs of cars over a period of time could arouse suspicion.

- **Rendering hidden data useless**

Once a file is identified as possibly containing hidden data, one can either attempt to recover the information if the algorithm is known, or to destroy the data without affecting the quality of the original file. An altered bitmap converted to JPEG would compress the file and remove the unnecessary bits of information, therefore removing any hidden data.

Converting to any other format may not necessarily cause the image to lose information, but would change the bit composition of the data, making any hidden data unreadable.

3.4 Steganalysis

There are two stages involved in breaking a steganographic system: detecting that Steganography has been used and reading the embedded message. The goal of a steganalyst is to detect stego message, read the embedded message and prove that the message has been embedded by a third party. Detection involves observing relationships between combinations of cover, message, stegomedia and Steganographic tools. This can be achieved by passive observation. Whether a message has been in an image or not, the image could be manipulated to destroy a possible hidden message.

Attacks on Steganography can involve detection and/or destruction of the embedded message. A ‘stego-only’ attack is when only the stego image is available to be analyzed. A ‘known-cover’ attack is one where the original cover image is also available. It involves comparing the original cover image with the stego image. Hiding information results in alterations to the properties of a cover image, this may result in some sort of degradation to the cover image. Original images and stego images can be analyzed by looking at color composition. Luminance and pixel relationship and unusual characteristics can be detected.

• Salt and Pepper Noise Attack

As the name suggests salt(white) and pepper(black) this type of noise introduces black & white dots on the image i.e. any pixel in our image will be modified to gray level near to 0 or $(2^N)-1$ in case of an N bit image. Our technique gives good results against Salt & pepper noise.

• Cropping

If some part of the image is cut it is called cropping. Cropping may cause distortion of the hidden message in the stego-image. Our technique provides reasonable resistance against cropping.

• Compression

Compression of the image causes distortion of the hidden message. If the message is distorted and compressed hidden message cannot be extracted Our technique extracts the hidden message accurately when the Stego image is PNG compressed.

4 Results

Table 1. Analytical Results

Results with AES as encryption technique			
Covers Selected	Execution Time (Seconds)	Accuracy in % .	MSE
Blue Cover, Multicolor Cover, Baboon Cover	49.0329 44.8796 45.6082	with Blue 100 with Red 100 with Green 100 with RGB 100	with Blue 0 with Red 0 with Green 0 with RGB 0
Results after cropping 30 rows and 150 columns from stego images			
With Multicolor Cover	42.5569	with Blue 100 with Red 5.72 with Green 17.3 with RGB 0.173	with Blue 0 with Red 1.708e+004 with Green 1.568e+004 with RGB 2.056e+004
Results after cropping 50 rows and 50 columns from stego images			
With Baboon Cover	42.3316	with Blue 100 with Red 0.347 with Green 100 with RGB 33.68	with Blue 0 with Red 3.935e+003 with Green 0 with RGB 2.823e+003

Table 1. (continued)

Results with AES encryption technique and Salt and Pepper noise 0.01			
Covers Selected	Execution Time (Seconds)	Accuracy in % .	MSE
With Blue Cover	45.1583	with Blue 55.5 with Red 52.77 with Green 58.3 with RGB 55.55	with Blue 2.02e+003 with Red 2.22e+003 with Green 1.69e+003 with RGB 1.76e+003
With Multicolor Cover	44.6841	with Blue 55.9 with Red 58.33 with Green 47.4 with RGB 55.55	with Blue 2.91e+003 with Red 2.81e+003 with Green 3.82e+003 with RGB 2.94e+003

Table 2. Pictorial Results





































I) AES as encryption technique			
Input Image	Encrypted Cipher Image	Cover image	Stego Image
			
Retrieved Cipher From Red Comp. 	Retrieved Cipher From Blue Comp. 	Retrieved Cipher From Green Comp. 	Retrieved Cipher From RGB Comp. 
Decrypted Image From Red Comp. 	Decrypted Image From Blue Comp. 	Decrypted Image From Green Comp. 	Decrypted Image From RGB Comp. 

Table 2. (continued)

II) Cropping 50 rows and 50 columns from stego			
InputImage 	Encrypted CipherImage 	Cover image 	Stego Image 
Retrieved Cipher From Red 	Retrieved Cipher From Blue Comp. 	Retrieved Cipher FromGreen 	Retrieved Cipher FromRGB 
Decrypted Image From Red . 	Decrypted Image From Blue Comp. 	Decrypted Image FromGreen 	Decrypted Image FromRGB 
III) With Salt and Pepper noise = 0.01			
Input Image 	Encrypted Cipher Image 	Cover image 	Stego Image 
Retrieved Cipher From Red Comp. 	Retrieved Cipher From Blue Comp. 	Retrieved Cipher From Green Comp. 	Retrieved Cipher From RGB Comp. 
Decrypted Image From Red Comp. 	Decrypted Image From Blue Comp. 	Decrypted Image From Green Comp. 	Decrypted Image From RGB Comp. 

5 Conclusion

We have implemented Palette Based Technique for Image Steganography and we have drawn following conclusions from the implementation. Our technique, when uses Advanced Encryption Standard as the encryption algorithm, provides very high security in a shared desktop environment. This is because AES is one of the strongest encryption standard available. So even if the secret message gets detected, it would be very difficult to decrypt it. However our technique uses AES which is not very

robust against the various attacks. This is because AES is a Block Cipher and change in one of the input bits causes the entire output to change. To make our technique resistant against Salt and Pepper noise, PNG Compression and Cropping the encryption algorithm 'Variation of Rail Fence' could be a good solution.

Also as the no of colors in the palette is reduced, the image quality also decreases. But addition of new color values to the palette does not affect the quality of the image as long as it doesn't change the original colors in the palette. This concept is used to hide text in the newly added color values of the palette which provides the identical stego image as that of cover image since text is hidden in palette colors and not in image pixel values.

The methodology stretches the palette of an image by adding two similar, but not identical color values for each color in the palette until it reaches maximum length of color palette or until colors are added to each original color in the palette.

Further Research should go towards enhancing the embedding process & adding extra functionalities such as hiding the secret data in different parts of cover image. Also trying to make the technique more robust against Zooming, JPEG Compression & Gaussian Noise should be included in further work.

References

1. Neeta, D., Snehal, K., Jacobs, D.: Implementation of LSB Steganography and Its Evaluation for Various Bits. In: 1st International Conference on Digital Information Management, pp. 173–178 (2006)
2. Hedieh S., Jamzad, M.: Cover Selection Steganography Method Based on Similarity of Image Blocks In : IEEE 8th International Conference Computer and Information Technology Workshops, pp. 379 – 384 (2008)
3. Samaratunge, S.G.K.D.N.: New Steganography Technique for Palette Based Images. In: Second International Conference on Industrial and Information Systems, University of Colombo School of Computing (UCSC), Sri Lanka (2007)
4. Cheddad A., Condell, J., Curran, K., McKeivitt P.: Enhancing Steganography in Digital Images. In : Canadian Conference Computer and Robot Vision, pp. 326 – 332 (2008)
5. Arjun, N.S., Negi, A.: A High Embedding Capacity Approach to Adaptive Steganography. In: 1st International Conference Digital Information Management, pp. 525–530 (2006)

Bilingual Malayalam – English OCR System Using Singular Values and Frequency Capture Approach

Bindu A. Thomas¹ and C.R. Venugopal²

¹ Vidya Vikas Institute of Engineering and Technology

² S.J. College of Engineering, Electronics and Communication Department, Mysore, India
binduthomas25@yahoo.co.in, venu713@gmail.com

Abstract. In India, bilingual documentation is very common especially in government forms and formats, technical documents, reports, postal documents, railways reservation forms etc., Printed documents having a single Indian language often contain English words and numerals since English is considered as a link language in India. The proposed system is designed to recognize bilingual script having Malayalam and English interspersed at word-level. This problem was considered as it is more realistic. Here, a combined database approach is employed, the scripts involved are treated alike and hence a single OCR is sufficient for recognition of bilingual script. The inherent advantage of the system is that the recognition of Malayalam, English words and numerals present in a bilingual document was achieved without performing script identification initially. This method avoids the script identification process which is computationally expensive. The proposed system achieves a recognition rate of 97.5% and 98.5 % for the two feature extraction approaches respectively.

Keywords: Bilingual character recognition system, Segmentation, frequency Capture features, Singular Value decomposition approach.

1 Introduction

The development of multi-lingual document image analysis systems has become an important task with wide range of applications. The diverse linguistic scenario in India places a tremendous demand on the computation industry to develop multilingual systems for machine translation, information retrieval, man-machine interfaces and so on. Consequently, requirements that arise include robust character recognition systems to convert document images in multiple Indian languages into text. The factors that make the task of the OCR designers even more challenging are the large database and the complex structure of Indian script and further the different combinations of these characters makes it even more difficult.

An extensive survey was conducted in this direction and the reporting include, a system for Han or Latin based script separation using fractal-based texture features developed by Spitz [1]. Tan [2, 3] developed an automatic method for identification of Chinese, English, Greek, Russian, Malayalam and Persian text. Among Indian scripts, Pal, *et. al.* [4, 5] proposed a generalized scheme for line-wise script identification from

a single document containing all Indian scripts. Dhanya, *et. al.*, [6] used Linear Support Vector Machine (LSVM), K-NN and Neural Network (NN) classifiers on Gabor-Based and Zoning features to classify Tamil and English scripts. Zhou, *et. al.*, [7] proposed a Bangla/English script identification scheme using connected component analysis. Recently, Jaeger, *et. al.*, [8] used K-NN, SVM, weighted Euclidean distance, and Gaussian mixture model to identify English from Arabic, Chinese, Korean and Hindi scripts. Jawahar, *et. al.*, [9] propose a Bilingual OCR for Hindi-Telugu documents based on Principal Component Analysis followed by support vector classification. The conclusion of this survey on multi-script OCR is that majority of the researchers have incorporated script identification as the first level in the design of a multi-script OCR. Most common assumption made was that the text document has same script at a paragraph/block or line level. But in our design a more realistic assumption is made; like in most practical situations the bilingual document has script changing at the word level.

2 Properties of the Participating Scripts

The knowledge of the characteristic properties of the scripts participating in the bilingual OCR plays a very important role. Thus knowing the size of the database (character set) and its structural complexity which is the inherent nature of most Indian script and especially south Indian language like Malayalam helps in the design of the bilingual Malayalam - English OCR.

Malayalam Script has 15 vowels and 36 consonants known as basic characters and these combine to take new shapes known as extended characters. Most of the characters in this script are very curly in structure added to their varying sizes mostly in terms of the width of the characters. The distinct property of the English characters and numerals is the existence of the vertical strokes. From the experiment, it is observed that vertical strokes are more dominant as compared to the horizontal strokes or closed loops (B, D, O, P, Q, a, b, d, etc.) and in case of numerals like 0, 6, 8, and 9. Some of the similarities between the two scripts are that both Malayalam and English script have unconnected character segments. Both the scripts are non-cursive and hence character segmentation is achieved with 100% accuracy.

3 Bilingual Malayalam- English Script Recognition System

The various stages involved in the development of the proposed system include image acquisition, preprocessing, segmentation, normalization, feature extraction and classification. A printed document having Malayalam text interleaved with English words and numerals is scanned on a flatbed scanner at 300 dpi for digitization. This digitized image is preprocessed for removal of background noise and the grey scale image is binarized using global thresholding technique. After preprocessing, the text image needs to undergo segmentation and normalization. It is assumed that the numerals will not be mixed with characters of both the scripts and hence after word-level segmentation the string is either numeral of characters. The proposed system does not use any separate script identification routine.

4 Bilingual Script Segmentation Technique

The similarity between the two participating scripts is the space between the characters within a word. The first stage of segmentation is based on the classical projection profiling techniques. Horizontal projection technique is applied to achieve line segmentation of printed bilingual documents. After line segmentation the obtained line segments undergo word segmentation using vertical profiling. These segmented words undergo one more level of vertical profiling wherein the characters, sub characters, numerals are segmented out. A section of a sample bilingual document and its vertical projection profile is shown in Figure 1.

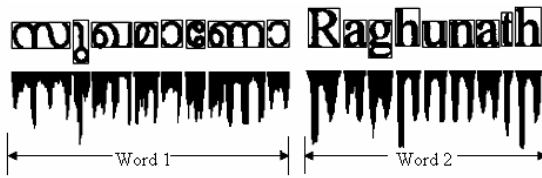


Fig. 1. Segmentation result at word, character, subcharacter level and their vertical projection profiles

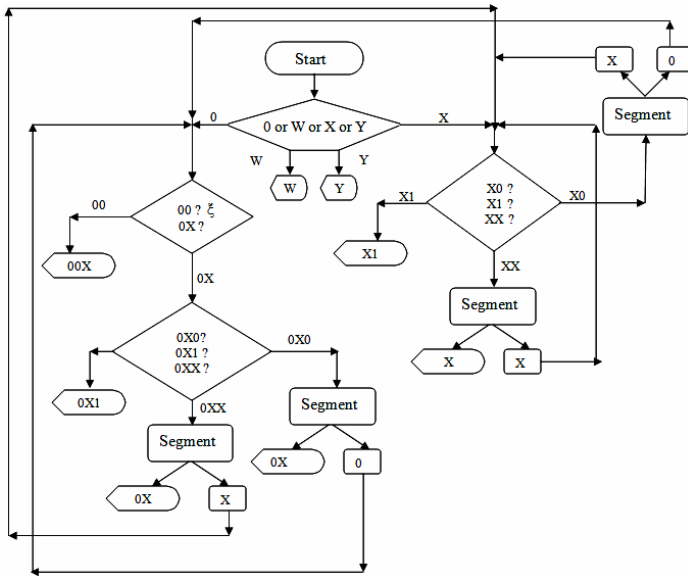


Fig. 2. OCR Segmentation Logic Flow Chart for Bilingual Script

The classical segmentation technique based on profiling returns the English characters, numerals and some basic characters of the Malayalam script but fail to segment the complete character of Malayalam script correctly because of the equal space between the characters of a word and the sub characters of a character within a word.

The string of segmented characters is thus fed as input to the next level of segmentation based on our novel segmentation approach. Figure 2 shows the segmentation flow chart. Here W represents English character (upper or lower case) or numerals and X represents a Malayalam consonant/ conjunct while Y represents a Malayalam vowel. The valid character sequences are of the form W, Y, X, X1, 0X, 0X1 and 00X. The logic used behind the choice of search space for the classifier is based on the sequence of arrangement of the segments of the character. For the sake of comparison the features are selected based on two approaches viz. the frequency capture and singular value decomposition approach. Further, these features are used for classification using nearest neighbor classifier, based on Euclidean distance. The inclusion of the 52 characters and 10 numerals of the English script in addition to the already existing large Malayalam character set database does not really have an effect on the classification task or the recognition accuracy. Both English and Malayalam characters are handled equally irrespective of the script they belong to.

5 Feature Extraction Approaches

The purpose of feature extraction is to identify appropriate measures to characterize the images with minimum number of features that are effective in discriminating pattern classes.

5.1 Frequency Capture Approach

This process in principle captures the frequency of transitions along each row. The feature vector, $x \in \mathbb{R}^m$ of the matrix $A \in \mathbb{R}^{m \times n}$ in this approach is defined by

$$x_i = \sum_{j=1}^n |a_{i,j+1} - a_{i,j}| \tag{1}$$

This captures features of characters with multiple loops a distinctive feature of Malayalam characters.

5.2 Singular Value Decomposition Approach

In this approach too in some sense, the row information is captured during the process of arriving at the singular values. Now, if $A \in \mathbb{R}^{m \times n}$, then \exists Orthogonal matrix

$$U \in \mathbb{R}^{m \times m} \text{ and } V \in \mathbb{R}^{n \times n} \ni U^T A V = \text{diag}(\sigma_1, \dots, \sigma_s) \in \mathbb{R}^{m \times n} \tag{2}$$

where $s = \min\{m, n\}$ and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_s \geq 0$.

For the extended Malayalam character set considered here, $s = m$ always for all images. These singular values are used to define a feature vector $\hat{x} \in \mathbb{R}^m$ defined by

$$\hat{x}_i = \sigma_i, \text{ where } i = 1, 2, \dots, m. \tag{3}$$

Let $\sigma_i = \{\sigma_1, \sigma_2, \dots, \sigma_p, \sigma_{p+1}, \dots, \sigma_m\}$. In our examples it is seen that $\sigma_i \gg \sigma_j$ for $i = 1, 2, \dots, p$ and $j = p + 1, \dots, m$, implying that there are few p dominant singular values. It is found experimentally that these p dominant singular values are sufficient to characterize the image matrix. The feature vector defined in Equation (3) now becomes

$$x_i = \sigma_i, \text{ where } i = 1, 2, \dots, p. \tag{4}$$

This definition of dominant singular values is further justified by the fact that $\| \|x\| - \|\hat{x}\| \| \leq \xi$, $\hat{x} \in \mathbb{R}^m$ and $x \in \mathbb{R}^p$, ξ being small, implying that there is insignificant contribution to the feature of the image from the $(m-p)$ singular values not being considered. The value of p is chosen by defining a value of ξ for reliable and robust classification.

6 Practical Example Used for Experimentation

Some of the practical and realistic examples used for experimentation are postal address, matrimonial, classifieds *etc.*,

Postal address with pin code written in English:

നിജിൽ കെ	Nijil.ke
കേളോത്ത് ഹൗസ്	kelloth House
ആഡൂർ	Aadoor
P O കെരള	P O Kerala
PIN 670621	PIN 670621
കെരള കണ്ണൂർ	Kerala, Kannur

7 Evaluation of OCR System for Bilingual Script at Word-Level

The suggested methods for bilingual OCR system are tested and evaluated using sample documents and the details of the investigation is presented in this section. The testing was performed both at word-level as well as character-level in order to evaluate the performance of the system. The data for testing the developed system was obtained by scanning 175 different documents from various novels, text books, magazine, *etc.* The scanned document images were subjected to pre-processing followed by word-level segmentation. After segmenting the words from the documents a test set of 5000 samples of Malayalam words and 3000 samples of English words inclusive of English numerals (800) were formed totaling 8000 words (which were different from training set). These words both Malayalam and English were so carefully chosen so as to include all combinations of the characters within the entire character set of both Malayalam and English scripts. Table 1 shows the results obtained for the feature extraction techniques evaluated at word-level.

Table 1. Recognition results for character -level Malayalam, English characters and English numerals

Word-level Recognition Accuracy						
Sl.No.	Method	Script	Samples	Hit	Miss	Recognition
1.	Frequency Capture	Malayalam	5000	4871	129	97.42 %
		English	2200	2162	38	98.27 %
		English number	800	781	19	97.62 %
2.	SVD	Malayalam	5000	4928	72	98.56 %
		English	2200	2167	33	98.50 %
		English number	800	788	12	98.52 %

8 Conclusions

The SVD though computationally complex results in an recognition accuracy of 98.56% for Malayalam words, 98.50% for English words and 98.52% for English Numerals while the Frequency capture is of 97.42 %, 98.27 %, 97.62 % for Malayalam words, English words and numerals respectively.

References

1. Spitz, A.: Determination of the Script and Language Content of Document Images. IEEE Trans. PAMI 19(3), 235–245 (1997)
2. Tan, T.N.: Rotation Invariant Texture Features and Their Use in Automatic Script Identification. IEEE Trans. PAMI 20(7), 751–756 (1998)
3. Zhou, L., Lu, Y., Tan, C.L.: Bangla/English script identification based on analysis of connected component profiles. In: Bunke, H., Spitz, A.L. (eds.) DAS 2006. LNCS, vol. 3872, pp. 243–254. Springer, Heidelberg (2006)
4. Pal, U., Choudhuri, B.B.: Indian Script Character Recognition: A Survey. Pattern Recognition 37, 1887–1899 (2004)
5. Pal, U., Chaudhuri, B.B.: Automatic Separation of Different Script Lines from Indian Multi-script Documents. In: Proc. Indian Conference on Computer Vision, Graphics and Image Processing, pp. 141–146 (1998)
6. Dhanya, Ramakrishna, A.G., Pati, P.B.: Script identification in printed bilingual documents. Sadhana 27(part-1), 73–82 (2002)
7. Zhou, L., Lu, Y., Tan, C.L.: Bangla/English script identification based on analysis of connected component profiles. In: Bunke, H., Spitz, A.L. (eds.) DAS 2006. LNCS, vol. 3872, pp. 243–254. Springer, Heidelberg (2006)
8. Jaeger, S., Ma, H., Doermann, D.: Identifying Script on Word-Level with Informational Confidence. In: Proceedings of the 8th International Conference on Document Analysis and Recognition, pp. 416–420 (2005)
9. Jawahar, C.V., Pavan Kumar, M.N.S.S.K., Ravi Kiran, S.S.: A Bilingual OCR for Hindi-Telugu Documents and its Applications. In: International Conference on Document Analysis and Recognition (October 2003)

Minimization of Handoff Failure and Cellular Traffic by Introducing IEEE 802.11b WLAN Router in the Handoff Region

Tapas Jana¹, Joydeep Banerjee², Indranil Chakroborty², Tara Sankar Patra¹, Debabrata Sarddar², M.K. Naskar², and Utpal Biswas³

¹ Dept. of ECE, Netaji Subhash Eng. College, Kolkata-700152

tjanansec@gmail.com, patratarasankar@gmail.com

² Dept. of ETCE, Jadavpur University, Kolkata-700032

jogs.1989@rediffmail.com, riju2205@gmail.com,

dsarddar@rediffmail.com, mrinalnaskar@yahoo.co.in

³ Dept. of CSE, University of Kalyani, West Bengal Nadia-741235

Utpal01in@yahoo.com

Abstract. Handoff is an inherent drawback of mobile communication, especially in urban areas, due to the limited coverage of access points (APs) or base stations (BS). It is essential to minimize this delay to provide the user with seamless network coverage. Many people have applied efficient location management techniques in the literature of next generation wireless system (NGWS). However, seamless handoff management still remains an open matter of research. Here we propose to minimize the handoff failure probability by effectively placing a wireless local area network (WLAN) AP in the handoff region between two neighboring cells. The WLAN coverage, on one hand, provides an additional coverage in the low signal strength region, and on the other hand, relieves the congestion in the cellular network. Moreover, we perform the channel scanning within the WLAN coverage area, thus minimizing the handoff failure due to scanning delay.

Keywords: WLAN, IEEE 802.11, Handoff. Handoff latency.

1 Introduction

IEEE 802.11b standards have become increasingly popular and are experiencing a very fast growth upsurge. They are widely being deployed for variety of services as it is cheap, and allow anytime or anywhere access to network data. However, they suffer from limited coverage area problem [1] and it is necessary to use this technology in the most prudent manner.

A Wireless Local Area Network (WLAN) links two or more devices using some wireless distribution method (typically spread-spectrum), and usually provides a connection through an access point to the wider internet. This gives users the mobility to move around within a local coverage area and still be connected to the network [2].

Essentially, a WLAN is an extra backbone to the connection between a mobile client and an access point, which transmits and receives radio signals between them.

An access point can be either a main, relay or remote base station. A remote base station accepts connections from wireless clients and passes them to relay or main stations. Connections between "clients" are made using MAC addresses [2].

The entire handover procedure can be divided into scanning, authentication and re-association. In the first phase, the mobile node (MN) scans for AP's by either sending Probe Request messages or by listening for beacon message. After scanning all or specified number of channels, an AP is selected which has the frequency channel with the highest Received Signal Strength Indication (RSSI) and Carrier to Interference (CI) ratio. Then the selected AP exchanges authentication messages with the MN. Finally, if the AP authenticates the MN, it sends Re-association Request message to the new AP and handover gets accomplished. The handover region between two approximated hexagonal coverage area cells for mobile communication is shown in Fig 1 (figure taken from Mobile- man).

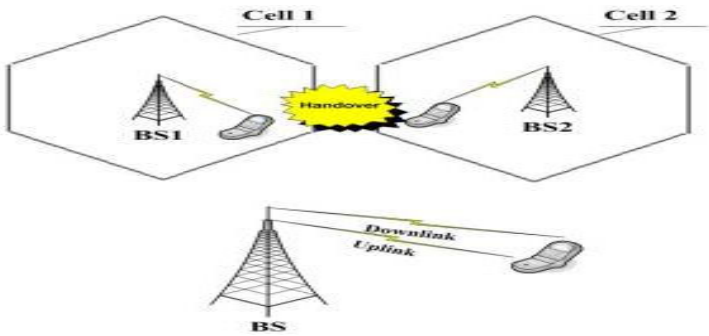


Fig. 1. The handover region between two approximated hexagonal coverage area cell

According to [3], 90% of the handoff delay comes from channel scanning. The range of scanning delay is given by the equation $N \times T_{min} \leq T_{scan} \leq N \times T_{max}$ Where N is the total number of channels according to the spectrum released by a country, T_{min} is Minimum Channel Time, T_{scan} is the total measured scanning delay, and T_{max} is Maximum Channel Time.

One of the most important reasons of handoff failure is the handoff latency caused by channel scanning and excess wireless traffic. Many measures have been taken in order to minimize handoff failure, but handoff failure is still an issue unsolved in the cellular world. Here we propose to minimize the handoff failure probability by effectively placing a WLAN AP in the handoff region between two neighboring cells. We also perform the channel scanning (required for horizontal handover between the two base stations) within the WLAN coverage area, thus minimizing the handoff failure due to scanning delay.

2 Related Works

In recent times, a large amount of research is done in improving the handoff technologies of cellular as well as IEEE 802.11 based networks. In the past few years, many methods based on neighbor graph [4] and geo-location on APs [2] has been proposed,

where the authors have proposed selective channel mechanisms. In [5] Li Jun Zhang et al. proposes a method to send probe requests to the APs one after the other and perform handoff immediately after any AP sends the response. This allows us to scan fewer channels. All these processes involve scanning of APs, it may be selective or all APs may be scanned. These methods are therefore time consuming as well as have a certain probability of handoff failure. In [6] and [7], authors use GPS based access point maps for handoff management. Handoff using received signal strength (RSS) of BS has been proposed previously. Using dynamic threshold value of RSS for handoff management for MNs of different velocities has been described in [8].

3 Proposed Work

We utilize the network coverage of a WLAN router for the purpose to reduce the handoff failure probability. If the traffic density is high then there will be a high handoff failure probability of the approaching MN. By integrating a WLAN with cellular networks, the traffic density of the cellular network (CN) is partially reduced, thereby minimizing the handoff failure probability to a great extent. In cellular network, the coverage area is approximated to be divided into a number of hexagonal cells. Let us consider two adjacent cells. We define threshold signal strength of a cell as the signal strength after which the handoff is initiated. We place a WLAN router between the threshold signals of either cell i.e. with the router being at the mid point of the line of the two AP's as shown in Fig 2.

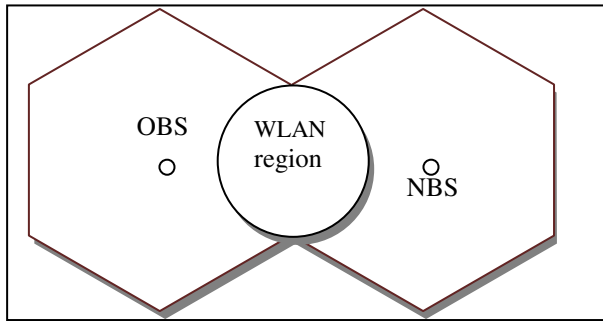


Fig. 2. The position of the WLAN in the handoff region

For WLAN the most prominent standard in use today is IEEE 802.11b which provides a bit rate of 11 Mbit/s and operates on 11 out of 14 distinct frequency channels. Path loss for a WLAN (PL) is given by $PL=L+10 \times Y \times \log(d) + s$ Here L is constant power loss, Y is the path loss exponent with values between 2 to 4, d represents the distance between the mobile node and WLAN AP and S represents shadow fading which follows Gaussian distribution. The Received Signal Strength (RSS) for WLAN (RSS_{WLAN}) in dBm is:

$$RSS_{WLAN} = PT - PL \tag{1}$$

Here PT is the power transmitted.

There are different wireless network standards under IEEE 802.11 workgroup and is shown below in Table 1:

Table 1. IEEE 802.11 standards and its specification

IEEE STANDARDS	802.11 protocol	a	b	g	n
	Frequency band(GHz)	4.915 – 5.825	2.412 – 2.484	2.412 – 2.484	4.915 – 5.825
	Bandwidth(MHz)	20	20	20	40
	Data Rate(Mbit/s)	6, 9, 12, 18, 24, 36, 48, 54	1, 2, 5.5, 11	1, 2, 6, 9, 12, 18, 24, 36, 48, 54	15, 30, 45, 60, 90, 120, 135, 150
	Approx indoor range(m)	35	38	38	70
	Approx outdoor range(m)	120	140	140	250

We will be working with IEEE 802.11 b standard WLAN AP. Received signal strength (RSS) for this standard incorporating the path loss from equation no. (1) is tabulated in Table 2 as follows:

Table 2. The RSS values for various position of MN for WLAN AP

Range	Received signal strength
50	-59.67
100	-68.70
150	-73.98
200	-77.73

3.1 Change of Base Station in Cellular Network and the WLAN Implementation

When the mobile node is certain to move into a particular base station, it starts the scanning process. The mobile node scans for the channels in the new base station, while under the coverage area of WLAN. The channel scanning process mostly contributes to handoff latency. The mobile node scans the channels which the NBS uses. Here, we reduce the number of base stations to be scanned to 1. Hence the number of channels to be scanned obviously becomes very low. This scanning process occurs under the network coverage of WLAN. Hence, there is minimum handoff failure probability, which occurs mainly due to scanning delay during a handoff process. As the scanning process terminates, the mobile node sends authentication requests and then the re-association requests which involves only two signals to be sent.

The proposed method for handoff is advantageous to normal handoff because of the fact that in this case, the mobile node in most of the cases is connected to the

network directly to the CN AP or to WLAN AP and hence it is expected to have a low probability of call dropping. In normal handoff, if the scanning process is time consuming due to high traffic density, the mobile node leaves the handoff region before establishing connection with NBS, resulting in handoff failure. In our proposed method, the mobile node after a specific threshold is under WLAN connection (if the traffic density permits) and even if it leaves the handoff region before establishing a connection with the NBS, it will be connected to WLAN until it gets connected with the NBS and hence there is minimum handoff failure in this case.

Here, we conduct the scanning process inside the WLAN coverage area, such that the scanning delay is completely eliminated from the handoff scheme as it no longer affects the handoff failure probability. Thus the Effective Handoff delay equals the summation of Authentication Delay and Re-association Delay.

3.2 Reduction of Traffic due to Introduction of WLAN

This handoff scheme allows a number of users to free the channels of the cellular network and avail for the WLAN coverage. This decreases traffic density of the existing APs.

If $E = \lambda h =$ total traffic, P_b is the probability of blocking, m is the number of resources such as servers or circuits in a group

$$P_b = B(E, m) = \frac{E^m / m!}{\sum_{i=0}^m E^i / i!} \tag{2}$$

Here we divide the average arrival rate λ into λ_{CN} and λ_{WLAN} where

$$\begin{aligned} \lambda &= \lambda_{CN} + \lambda_{WLAN} \\ E_{CN} &< E \\ \text{New probability of blocking} &< P_b \end{aligned} \tag{3}$$

Thus, with this paper we endeavor to reduce the traffic density and call blocking probability in the handoff region in a cellular network by introducing a WLAN AP inside the handoff region.

3.3 Assignment of WLAN Channel to Specific Users

We design an algorithm on filtering the number of users who are given the option of availability of WLAN service. We locate the position of the old base station (OBS) and mobile-node (MN) by GPS technology.

We assume (X, Y) are the co-ordinates of the present BS and (x_m, y_m) are the co-ordinates of the MN and r is the radial distance of the mobile node from the OBS.

Then we have the radial distance r of the MN from its present BS is given by the equation,

$$r = \sqrt{(X - x_m)^2 + (Y - y_m)^2} \tag{4}$$

* We define R as signal range of the directional antenna used by the OBS */

/* We define D_{WLAN} as diametric range of WLAN coverage area */

1. While base station connectivity not changed
2. If $r < (R - D_{WLAN})$,
 - 2.1 The MN does not take any action of vertical handover
3. Else
 - /* We define \dot{r} = rate of change of r */
 - /* Now \dot{r} increases or decreases only if r changes */
 - /* We define T as the time taken to make a connection with WLAN AP */
 - 3.1. If $\dot{r} * 2T < R - r$
 - 3.1.1. Then mobile node will not be able to avail the WLAN service and wait till the next iteration comes
 - 3.2. Else
 - 3.2.1. The mobile node tries to make a connection with the WLAN AP.
 - 3.2.2. The mobile nodes will be preferentially allocated channels by the WLAN AP in order of there $\dot{r} * 2T$ value.
 - 3.3. End
4. End
5. End

This algorithm provides a filter to the number of users availing the WLAN service to avoid unnecessary vertical handover.

4 Simulation Result

We made simulation of our proposed method using the algorithm in the subsection of the proposed work. For justifying the practicability of our method in real models we made an artificial environment considering a seven cell cluster, that is seven APs at the centre of seven closely packed cells whose hexagonal structure provides an approximation of its signal sweeping region, and we implemented Poison's Distribution Function for incorporation of memory less property in the generation of calls in the environment.

Depending upon the level of traffic impediments we noted down the corresponding call blocking probability, that is the number of calls that terminates during handoff to the total number of handoffs in that traffic condition. The traffic of the network is taken relative to the maximum traffic. The plot of network traffic verses the call blocking probability is shown in the Fig 3. We have also shown the handoff call blocking probability, that is the number of calls that terminates during handoff to the total number of handoffs in that traffic condition, in the same base parameter in Fig 6.

From Fig 3 it can be seen that the call blocking probability is below 0.1 for cases up to where seventy percent of the network is congested with a gradual increase, which is obvious for high congestion in the network. Fig 4 shows that handoff call blocking probability increases in an exponential manner with network traffic but we are able to restrict it to 25% of the maximum value which is an significant improve over the previous results. Thus our method effectively reduces the traffic blocking probability and also handoff call blocking probability.

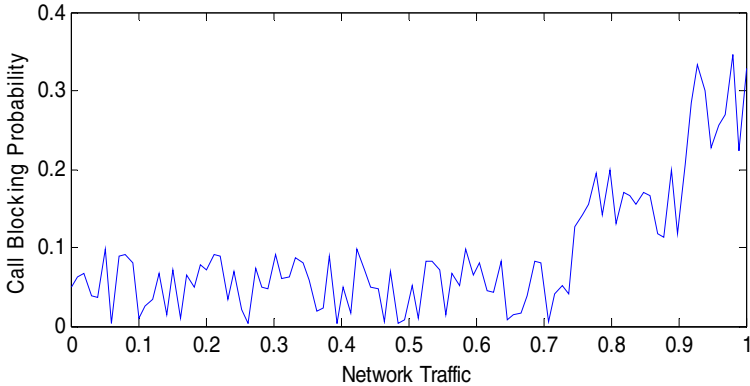


Fig. 3. Plot of network traffic versus the call blocking probability

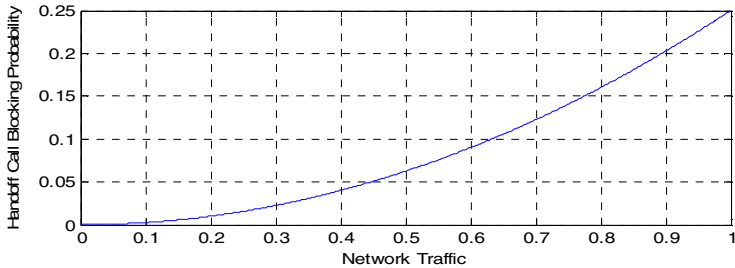


Fig. 4. Plot of network traffic versus the handoff call blocking probability

5 Conclusion

Thus by our proposed method, we can reduce handoff failure as well as handoff latency quite a remarkable amount as we can reduce the traffic in the cellular network by introducing a WLAN AP. The various advantages of incorporating the WLAN AP in the CN thus can be enlisted as follows. Firstly, this facility will relieve congestion on the GSM or UMTS spectrum by removing common types of calls and routing them to the operator via the relatively low cost Internet. Secondly, this scheme allows carriers to add coverage using low cost 802.11 access points. Subscribers enjoy seamless coverage. Thirdly, this handoff procedure cuts out the scanning delay from the handoff latency components by scanning the channels while in the WLAN coverage and finally, the handoff failure probability tends to zero. However, future works can be done on improving the traffic distribution between the CN and WLAN, so that handoff failure can be eliminated completely.

References

1. Sarddar, D., et al.: Minimization of Handoff Latency by Angular Displacement Method Using GPS Based Map. *IJCSI International Journal of Computer Science Issues* 7(3) (May 2010)
2. http://en.wikipedia.org/wiki/Wireless_LAN_Wikipedia (free encyclopedia)
3. Pesola, J., Pokanen, S.: Location-aided Handover in Heterogeneous Wireless Networks. In: *Proceedings of Mobile Location Workshop* (May 2003)
4. Stevens-Navarro, E., Pineda-Rico, U., Acosta-Elias, J.: Vertical Handover in beyond Third Generation (B3G) Wireless Networks. *International Journal of Future Generation Communication and Networking*, 51–58
5. Kim, H.-S., et al.: Selective Channel Scanning for Fast Handoff in Wireless-LAN Using Neighbor-graph. In: *International Technical Conference on Circuits/Systems Computers and Communication*, Japan (July 2004)
6. Dutta, A., Madhani, S., Chen, W.: GPS-IP based fast Handoff for Mobiles
7. Tseng, C.-C., Chi, K.-H., Hsieh, M.-D., Chang, H.-H.: Location-based Fast Handoff for 802.11 Networks. *IEEE Communications Letters* 9(4) (April 2005)
8. Mohanty, S., Akyildiz, I.F.: A Cross Layer (Layer 2+3) Handoff Management Protocol for Next-Generation Wireless Systems. *IEEE Transactions on Mobile Computing* 5(10) (October 2006)
9. Ayyapan, K., Dananjayan, P.: RSS Measurement for Vertical Handoff in Heterogeneous Network. *JATIT*, 989–994
10. Akyildiz, I., Xie, J., Mohanty, S.: A Survey of Mobility Management in Next-Generation All-IP-Based Wireless Systems. *IEEE Wireless Communications* 11, 16–28 (2004)
11. Gustafsson, E., Jonsson, A.: Always Best Connected. *IEEE Wireless Communications* 10, 49–55 (2003)

Design of GSM Based Auto-responder for Educational Institute

D.D. Vyas¹ and H.N. Pandya²

¹ Department of Electronics & Communication Engg. Darshan Inst. of Engg. & Tech.
Rajkot – 360 001, India

² Department of Electronics, Saurashtra University
Rajkot – 360 005, India
{vyasdd, hnpandya}@yahoo.com

Abstract. Remote access of data using off-the-shelf technologies like internet and GSM/GPRS, is one of the fast growing application areas in industries, customer services, banking, entertainment services etc. This paper presents design and implementation of an auto-responder for educational institute. The design is based on GSM Short Message Service and is used for remote access to data related to examination, results and student's attendance to be accessed by students or any stake holder, at their own wish. The complete application software, including database management, GSM communication and interactive GUI, is designed in MATLAB. The application provides a good example of the use of Information and Communication Technology (ICT) and illustrates how low cost off-the-shelf technologies like GSM can be meaningfully used for remote data access.

Keywords: GSM applications, Short Message Service (SMS), auto-responder, remote data access, MATLAB programming.

1 Introduction

Advancements in Information and Communication Technologies (ICT) provide easy access to data remotely using number of low cost off-the-shelf technologies like internet and mobile communication. Potential use of these technologies has been realized in almost all domains including industries, customer services, banking, entertainment services etc. and hence has emerged out as one of the fastest growing application areas [1], [2], [3]. Short Message Service (SMS) available on existing GSM network provides a low-cost solution for remote data access and can be easily used for monitoring and control functions just by adding small hardware. SMS is based on data packet switching provided on GSM networks and uses control channel allowing parallel voice communication. Use of SMS for remote data access has number of advantages. The biggest advantage is its simplicity in implementation and use. Further, they are very economic and are relatively fast. GSM encryption inherently provides data security in SMS transmission and features like delivery report; automatic retransmission in case of transmission failure, message broadcast etc. makes it more useful. The amount of data that can be sent via SMS is limited to 140 bytes which is sufficient in most applications [4].

In this paper design and implementation of a GSM-SMS based auto-responder for use in educational institute is presented. It is implemented to cater desired information related to examination, results and student's attendance to students or any stake holder, 24x7 on mobile phones through SMS request. This facility in particular helps parents to regularly keep track of punctuality and progress of their children from distant location and at their own wish. Most of the institutes maintain their website and hence one of the possible ways to keep parents posted about student's performance is by making related data available online or posting the same through e-mail. Though doing this is easy, but in Indian context, this does not serve the purpose much as most of the parents either does not or are not comfortable with the use of internet or sometimes even a computer.

2 System Architecture

The system is mainly composed of the central data server, GSM modem, remote GSM enabled mobile device and GSM network. The wireless remote communication between the data server and remote mobile device is realized by GSM network using GSM modem. The system architecture is shown in the Fig. 1.

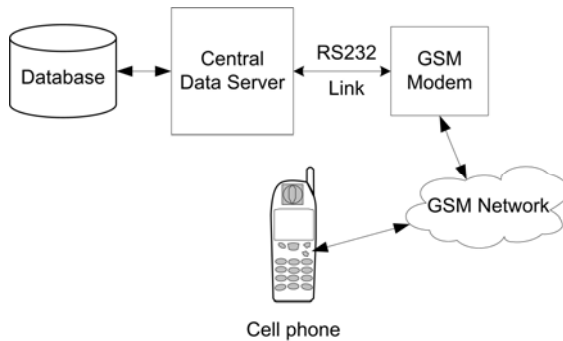


Fig. 1. System Architecture

The central data server can be implemented by using ordinary computer server, workstations and so on. Its main functions are:

- 1) to receive, classify and save the data related to attendance, examination and results given to it on day to day basis.
- 2) to manage database, extract required information and generate and print desired reports.
- 3) to respond to the request received from the remote user over GSM link in real-time, by identifying the requested service, checking its validity, extracting the relevant information and communicating it back.
- 4) to log the details of all communications done over the GSM network.

GSM modem is the key equipment responsible for establishing the link between the central server and GSM network. It acts as data terminal equipment that combines

traditional modem and GSM wireless mobile communication system. GSM modem used in this application is ST2133 from Scientech Technologies. The modem works in frequency bands of 900 MHz/1800 MHz, and supports baud rates maximum up to 14.4 kbps. ST2133 provides standard RS232 interface and hence can be directly connected to the serial interface of central data server. It provides the standard AT command support for user interface. The module provides fast, reliable and secure transmission of data, voice, short messaging and fax.

3 System Software

The complete system software has been designed using MATLAB. Though MATLAB is not specialized software for database management, it has wonderful ways of working with matrices (data structures), communicating over serial ports and designing front end GUI [5]. The block diagram of the system software is shown in Fig. 2. The software consists of four modules Database creation/amendment, Data collection and classification, Data analysis and report generation and GSM communication. All these modules are encapsulated in an easy-to-use GUI.

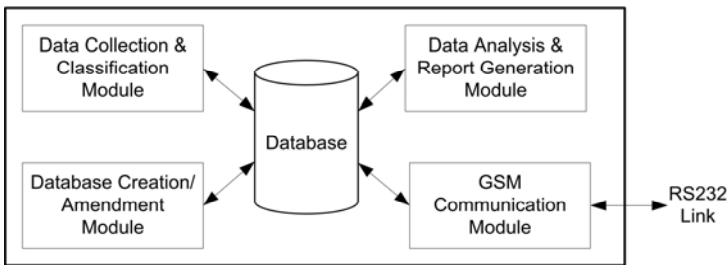


Fig. 2. Block diagram of System Software

3.1 Database Creation/Amendment Module

This module allows to create a fresh database at the beginning of an academic term and also allows to edit and amend the created data base as and when required. In MATLAB, MAT is the default data file format used to save the data on system disk. A typical example of a database structure for student's attendance created by this module is shown in Fig. 3. It is created for an institute with six branches of engineering and with two divisions (class) for each branch.

3.2 Data Collection and Classification Module

This module allows to feed data to the database and store them appropriately. For example, day-to-day cumulative attendance of lectures plus laboratory of each division given to this module is stored in attendance matrix against name and enrollment number of each student. The data to be stored can be given in form of a Microsoft excel file which will be extracted by the module and saved in the database in MAT data format.

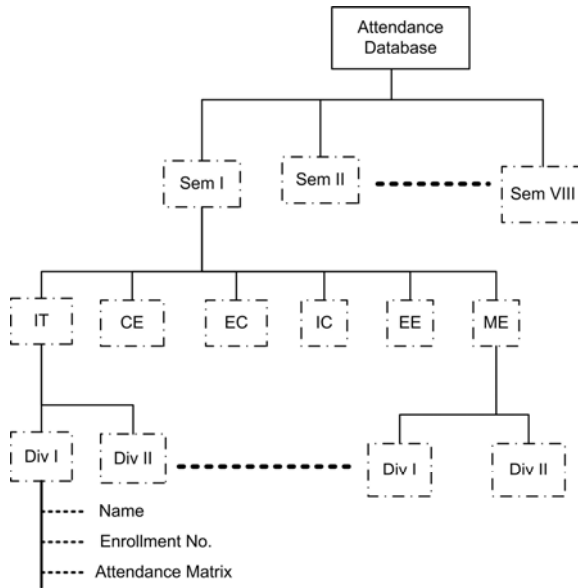


Fig. 3. Database structure for student's attendance

3.3 Data Analysis and Report Generation Module

This module helps to respond to database queries generated either from the data server console (for local reference) or from remote user. Module supports generation of detailed reports on basis of the local database queries, which can further be printed or saved for reference. For example it is possible to extract information about attendance of the whole class or any individual student between given period of time. The reports generated in response to the local database queries are in form of Microsoft Excel files. Standard text message format has been defined for database query from a remote user using SMS. This format is *<student's branch> <student's enrollment number> <requested information code>*. The requested information code can be A for query related to attendance, R for result and E for examination schedule. Thus a valid query can be an SMS like - CE 90010056 A. Messages only in this format and with predefined information code are considered as valid and responded.

3.4 GSM Communication Module

This module is responsible for receiving and responding to the SMS request received from the remote user over GSM network via GSM modem connected to the serial port of central data server. Communication with GSM modem can be enabled or disabled by the user. In a normal run the communication is always enabled. It is disabled only for a short time particularly when the database is being updated or edited. This module consists of code for following functions.

Initialization of serial port. This includes code for getting access of the serial port of central data server and setting put associated properties. These properties includes baud rate, data format (number of data bits, stop bit and parity), size of input and output buffers, mode of reading serial port (asynchronous or synchronous), timeout while attempting to read, message terminator etc.

Initialization of GSM modem. AT commands are used to control the functionality of GSM modem. Principal AT commands used in initialization of GSM modem are listed in Table 1. Two methods can be used to send and receive SMS messages, text mode and PDU (Protocol Description Unit) mode and both are based on AT command sets. In this application GSM modem is used in text mode and message reception procedure is set such that when ever a new message is received by GSM modem, it is stored in SIM memory and the location of this message is forwarded to central data server. After GSM modem is initialized serial port object is configured in an asynchronous mode so that the software will receive an interrupt when ever a request SMS is received.

Reading serial port and receiving SMS. This is a callback function executed when GSM modem sends notification of received SMS message to the serial port of central data server. The execution flow of this code section is shown as a flow chart in Fig. 4. The message is considered to be a valid request only if it is in the standard format as discussed before.

Writing serial port and sending SMS. This code fragment is called when a valid SMS request has to be replied. A text message is framed and is sent using AT+CMGS command. For example, a typical message replied against a query related to attendance will be like “Paras H. Joshi, Branch:CE, Sem:4, Attd:79.26%, upto:2/06/2010.” In case if the message received is in the correct format but the content is not valid then a reply message is sent informing the user to verify the content.

Deleting the served message from SIM memory. Once the reply SMS is successfully sent, the request message present in SIM memory is deleted using AT+CMGD by this code segment.

Table 1. Principal AT commands for GSM modem initialization

AT commands	Operation
AT	check successful connection between PC and GSM modem
AT+CPIN	detect valid SIM card
AT+CREG	check network registration of GSM modem
AT+CMGF	set preferred message format, text or PDU
AT+CNMI	select procedure for message reception from the network

Updating a service log file. This code segment is responsible for recording the details of each communication between the central data server and GSM modem into a log file. Microsoft Excel file is used as a log file in which information like requester's phone number, request date and time, branch and enrollment of student, requested information code. If the request has been replied then replied message and date & time of reply is also registered.

If communication between PC and GSM modem is disable because of some reasons, then every time it is enabled back, list of unread SMS is first read from the GSM modem using AT+CMGL command and request of all valid unread SMS is catered first.

4 Implementation and Results

The complete software has been designed using MATLAB versions 7 on Windows platform. GSM service of BSNL, India has been used. Information stored in service log file reveals that in a normal run, the overall response time to cater a SMS request is fewer than 10 sec.

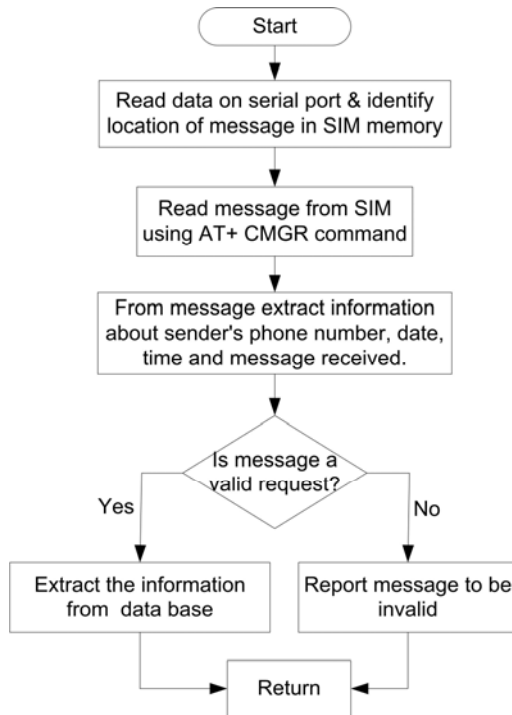


Fig. 4. Execution flow for reading serial port and receiving SMS

5 Conclusion

Design and implementation of a GSM-SMS based auto-responder for an educational institute has been discussed. The purpose of the system is to cater information related to examination, results and student's attendance to students or any stakeholder, 24x7 on mobile phones through SMS request. Different hardware and software unit of the system are described. The complete application software including database and the GUI has been designed using MATLAB. The results are presented which are quite satisfactory and the response received from the community in general is encouraging. The application furnishes a good paradigm for any remote data access application based on GSM communication network and PC based servers.

References

1. Yan, H., Pan, H.: Remote data monitoring system design based on GSM short message service. In: IEEE International Symposium on Industrial Electronics, pp. 364–369 (July 2009)
2. Chen, P., Jiang, X.: Design and implementation of remote monitoring system based on GSM. In: IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, pp. 678–681 (August 2009)
3. Wasi-ur-Rahman, M., Tanvir, M., Lutful, S.M.: Design of an intelligent SMS based remote metering system. In: IEEE International Conference on Information and Automation, pp. 1040–1043 (June 2009)
4. AT Commands Interface Guide, Wavecom website (2008), <http://www.wavecom.com>
5. MATLAB Documentation, MathWorks website (2010), <http://www.mathworks.com/access/helpdesk/help/techdoc>

Minimization of Handoff Latency by Area Comparison Method Using GPS Based Map

Tapas Jana¹, Joydeep Banerjee², Subhojit Banerjee², Souvik Kumar Mitra²,
Debabrata Sarddar², M.K. Naskar², and Utpal Biswas³

¹ Dept. of ECE, Netaji Subhash Engineering College, Kolkata-700152
tjanansec@gmail.com

² Dept. of ETCE, Jadavpur University, Kolkata-700032
jogs.1989@rediffmail.com, subhojit449@gmail.com,
souvikmitra.ju@gmail.com

dsarddar@rediffmail.com, mrinalnaskar@yahoo.co.in

³ Dept. of CSE, University of Kalyani, West Bengal Nadia-741235
Utpal01in@yahoo.com

Abstract. Handoff has become a drastic drawback in mobile communication system especially in urban areas owing to the limited coverage area of access points (AP). When a mobile node (MN) moves outside a range of its current access point it needs to perform a link layer handoff. This causes data loss and interruption in communication. In this paper we have proposed a process to minimize the handoff latency with the aid of GPS by minimizing the number of APs to be scanned by the MN during each handoff procedure. We have introduced a new algorithm based on the simplest concept of area. From GPS we can get an idea about the trajectory of motion of the MN and also the position of the neighbor APs. So taking these two factors into consideration we can reduce the scanning delay of the handoff to a great extent and thereby reducing the entire handoff latency.

Keywords: We IEEE 802.11, GPS (Global Positioning System), MN (mobile node), trajectory of MN, handoff latency, Area Comparison.

1 Introduction

IEEE 802.11 based wireless LANs have seen a very fast growth in the last few years and Voice over IP (VoIP) is one of the most promising services to be used in mobile devices over wireless networks. The main concern regarding wireless network technology is the handoff management. Complete handoff procedure can be divided into three distinct phases—scanning, authentication and re-association. The overall delay during handoff is sum of these three delays, and 90% of it is contributed by scanning delay. Scanning delay is calculated as

$$N \times T_{min} \leq T_s \leq N \times T_{max} \quad (1)$$

Where ' N ' is the total number of channels, T_{min} is Minimum channel time and T_{max} is Maximum channel time.

In this paper we propose a GPS based handoff mechanism. Each MN is equipped with a GPS receiver and sends its current position to its GPS server. This position is updated time to time. Taking the current co-ordinate of the MN and the co-ordinates

of the six vertices of the current AP, an algorithm based on area computation and comparison is designed in such a way that the server determines in advance the MN's next point of attachment and initiates handoff with the closest and the best AP. By this process, the delay incurred during scanning phase can be reduced. This paper is organized as follows First we present a brief idea of IEEE 802.11 standard handoff procedure, GPS in handover management and related works, followed by our proposed work. Finally we conclude with the simulation results and a brief conclusion.

2 Related Works

In [1] improvement of handoff latency by using neighbor graph, mobility graph and non-overlap graph was proposed. In [3] neighbor graph cache mechanism was proposed. In [2] the idea of selective scanning and caching mechanism was proposed and the idea of background scanning, pre-authentication and server based restricted channel set was given. Beside that the concept of implementing a hysteresis constant in addition to RSSI threshold was introduced, which is a very important tool to avoid the "togglng effect". In [4] some GPS based handoff technique is proposed. In [7], S. Kyriazakos et al. proposed an algorithm to resolve the well-known ping-pong and faraway cell effects using the MN's movement and its velocity. J Persola et al in [5] present a location assisted algorithm to manage handover between WLAN and GPRS networks.

3 Proposed Work

For any mathematical or practical purpose the coverage area of an AP is considered to be hexagonal with the AP situated at the centre of the hexagon and this hexagonal cell is surrounded by six other similar hexagonal cells (considering the antenna strength of each of these cells to be equal). These individual cells form the seven cell cluster as shown in Fig 1. Here we will consider the N-S and the E-W directions as the two directional axes perpendicular to each other and compare it with the x-y axes of the rectangular co-ordinate system.

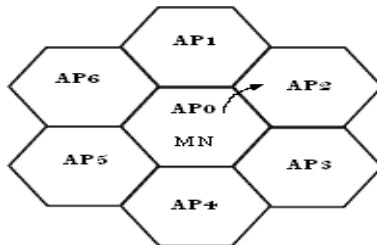


Fig. 1. A seven cell cluster with hexagonal approximation

Now, from Fig 2 we get the co-ordinates of the six vertices of the hexagon $P_1P_2P_3P_4P_5P_6$ as $P_i = (x_i, y_i)$, for $i = [1, 6]$. With the help of these six co-ordinates the

equation of the six edges of the hexagon are found out as follows $(y - y_i)/(x - x_i) = (y_i - y_{i+1})/(x_i - x_{i+1})$, which on simplification yields a linear two variable equation of form

$$a \times x + b \times y = 1 \tag{2}$$

Where, $a = ((y_{i+1} - y_i)/(x_{i+1}y_i - x_iy_{i+1}))$, and, $b = ((x_{i+1} - x_i)/(x_{i+1}y_i - x_iy_{i+1}))$. As soon as the MN enters a hexagon, the current co-ordinate of the MN is obtained with the aid of GPS. Let the current position be given by (p, q) . Taking it as the origin we obtain two straight lines $x=p$ (parallel to the N-S direction) $y=q$ (parallel to the E-W direction). This is shown in Fig 2. From equation (2) we can write $y=m_1x+c_1$ where $m_1 = ((y_i - y_{i+1})/(x_i - x_{i+1}))$, and $c_1 = ((x_{i+1}y_i - x_iy_{i+1})/(x_{i+1} - x_i))$, and $x=m_2y+c_2$, where $m_2 = ((x_{i+1} - x_i)/(y_{i+1} - y_i))$, and $c_2 = ((x_iy_{i+1} - x_{i+1}y_i)/(y_{i+1} - y_i))$ By putting $x=p$ and $y=q$ in the above two obtained equations respectively six set of co-ordinates will be obtained for each of the two equations. But we need only the four points which lie on the perimeter of the hexagon.

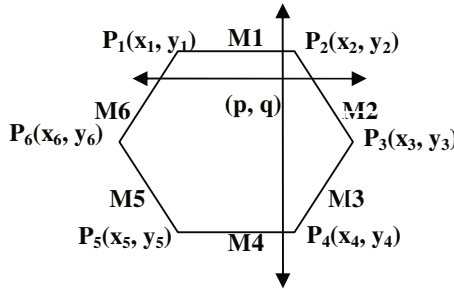


Fig. 2. The co-ordinates of the vertices of the approximated hexagonal AP

For the selection of these four points the following algorithm is followed:

1. The six vertices of the hexagon are taken and two consecutive vertices are taken at a time.
2. The 12 points obtained are taken each at a time.
3. It is compared if the co-ordinates of the point lie within the co-ordinates of the two vertices.
4. For e.g if the co-ordinate of the point be (p_i, q_i) , then it is checked if :

$x_i \leq p_i \leq x_{i+1}$	}
$y_i \leq q_i \leq y_{i+1}$	
5. If it lies then, that point will be stored ;otherwise it will be deleted.

Following the above algorithm the four points will be obtained. Then the area of the four parts intercepted between the two directional axes and the hexagon is found out. The co-ordinates that is required for the computation of each area is determined as follows:

1. Start with any one of the six vertices of the hexagon.
2. Observe if the vertex has been utilized to save any of the four points obtained from the previous algorithm.

3. If the vertex has such points on both sides (with respect to the adjacent vertices), then close the loop by taking the co-ordinates of the vertex, the two points and the current location of the MN.
4. If the vertex has one such point, then the search operation is shifted to the vertex on that side of the current vertex in which none of the four points were found. Then the same algorithm is carried in a recursive manner until another point of the four points is obtained.
5. The area is calculated by taking the co-ordinates of the two points, the intermediate vertices and the current location of the MN.
6. When the points are selected the corresponding areas are calculated using the formula as given by $0.5 * [\sum (x_i y_{i+1} - x_{i+1} y_i) + (x_n y_1 - x_1 y_n)]$, where $i = [1, n-1]$, where n is the number of points required to specify an area.
7. Four such areas are calculated and compared to find the least among them. Then it becomes obvious that the handoff procedure will occur through that part of the cell-boundary that constitutes a part of the perimeter of the minimum area calculated.

To predict the possible direction of approach of the MN the following algorithm has to be followed:

1. The slopes of the six edges of the hexagon are calculated in advance whenever an MN enters a BSS. $M_i = (y_i - y_{i+1}) / (x_i - x_{i+1})$, for $n = [1, 7]$ and $i = n \bmod 6$ when $n=1, 2, 3, 4, 5, 7$, and $=6$ when $n=6$.
2. The slopes of the sides of the four intercepted areas are also found out and then compared with the above equation.
3. The slope of the side of the hexagon, as shown in Figure 2, with which the slope of any side of the area matches determines the predicted target AP of the MN. Thereby the MN will have to scan the channels of only that AP and thereby the handoff latency will be reduced in total.

On the basis of Figure 1 and Figure 2 the predicted target APs are tabulated below:

Table 1. Prediction table

SLOPE	TARGET AP
M1	AP1
M2	AP2
M3	AP3
M4	AP4
M5	AP5
M6	AP6

However two situations may arise. Firstly, the smallest area may contain parts of two sides of the hexagonal cells. Secondly, the smallest area may contain parts of only one edge of the hexagonal cell.

The GPS server chooses the new AP either from two options (as in first situation) or from one option (as in second situation). The server sends a handoff initiate (HI) message to the MN and provides parameters like the target AP's IEEE 802.11 channels, SSID (service set identifier).

This entire handoff process starts after the signal strength received by the MN becomes less than a certain threshold value which depends upon the antenna strength of the AP. But the signal strength of the APs changes rapidly with space and time. It might happen that just after handoff the signal strength of the old AP is better than the current AP and thus the station initiates a handoff with the destination as the old AP. This effect is called the “togglng effect”. To avoid this togglng we have added a mechanism of hysteresis [4] i.e. the signal strength of the new AP must be better than the old AP by at least a hysteresis constant. Thus unnecessary handoffs can be reduced with the use of hysteresis. The parameters RSSI threshold and hysteresis constant are configurable and depends on the power of the transmitting signal.

As soon as the RSSI received by the MN goes below the threshold, then the minimum area is calculated, consequently the best AP is selected and the scanning phase starts followed by the authentication phase. Just at the moment the signal strength of the target AP becomes more than the old AP by the hysteresis constant then the process of re-association starts – thereby completing the handoff mechanism.

4 Simulation and Results

We have made a sample run of our algorithm to test the functionality of it. The coverage region of the AP is taken as the regular hexagon of side equal to 200m (which satisfies the topological conditions of an AP in urban areas). At the end of the algorithm we compare the prediction of our algorithm with the actual result and thereby justify the appropriateness of our algorithm. All the co-ordinates are in meters and are measured in reference to the present AP whose co-ordinates are (250,250). The initial speed of MN at the origin of call was taken as 20 m/s, and after the algorithm the average speed was recorded at 19.35 m/s and the speed variation is shown in Fig 3. The average handoff delay time was taken as 50 ms.

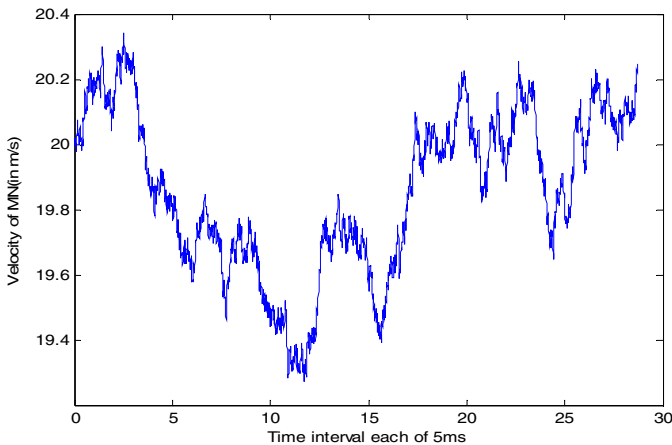


Fig. 3. The velocity of mobile node at different time instant

As proposed when the RSSI falls below a certain threshold (here we have considered it to be equal to 20 dB), the algorithm stops and the scanning starts. But the final handoff procedure is completed only when the hysteresis constant is realized.

In Fig 4 we have considered a parabolic trajectory of an MN. Comparing with Figure 1 we can see that 11 positions of the MN are taken into consideration, and corresponding to each point the probable APs are predicted.

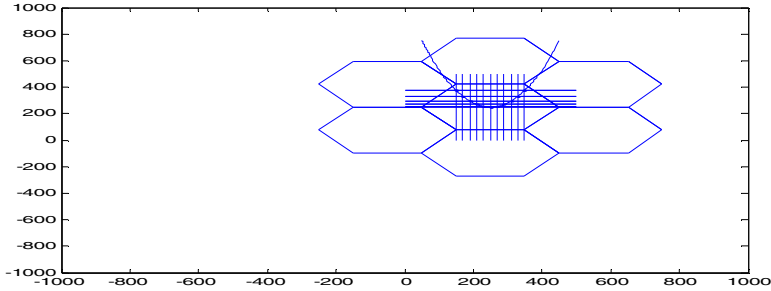


Fig. 4. Trajectory of MN

When the conditions are met our algorithm predicts the target AP to be AP2, which is true as can be viewed from Figure 4. Now in Figure 5 we have created a totally random trajectory of MN based on real world consideration.

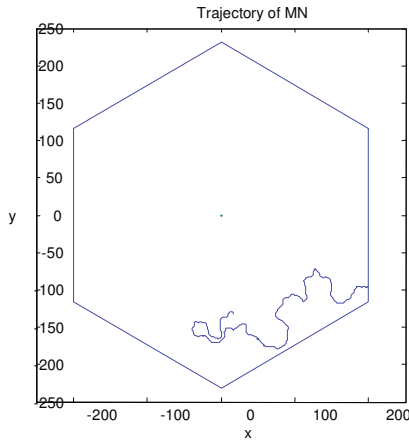


Fig. 5. Trajectory of the MN

Here a situation arises when the received RSSI goes below the threshold value, but as the hysteresis condition is not realized within the maximum scanning time so handoff does not occur to the AP3, and thereby the probability of false handoff is also reduced by our algorithm. And in this case handoff occurs finally with AP2.

5 Conclusion

By this area comparison method, we can reduce the handoff latency to a great deal as we can get a clear idea as to which channel to scan for a particular MN. The selection of the most potential AP by the MN effectively reduces the scanning delay as the number of channels scanned will be lower. Future work may include improvement in the cut-off condition after which the scanning procedure is initiated.

References

1. Shin, M., Mitra, A., Arbaugh, W.A.: Improving the latency of 802.11 handoffs using neighbour graphs. In: Proceedings of the 2nd International Conference on Mobile Systems, Applications and Services, Mobisys 2004, pp. 70–83. ACM Press, New York (2004)
2. Mishra, A., Shin, M., Arbaugh, W.A.: An empirical analysis of the IEEE 802.11 MAC layer handoff process. *SIGCOMM Comput. Common. Rev.* 33(2), 93–102 (2003)
3. Li, C.-S., et al.: A neighbor caching mechanism for handoff in IEEE 802.11 Wireless networks, March 20. Springer, Heidelberg (March 20, 2008), doi:10.1007/s11227-008-0175-3
4. Sarddar, D., et al.: Minimization of handoff latency by co-ordinate evaluation method using GPS based map. *International Journal of VLSI Design & Communication Systems* 1(2) (June 2010)
5. Pesola, J., Pokanen, S.: Location-aided Handover in Heterogeneous Wireless Networks. In: Proceedings of Mobile Location Workshop (May 2003)
6. Gast, M.S.: 802.11 wireless networks, the definitive guide, pp. 114–137. O'Reilly & associates, USA (2002)
7. Kyriazakos, S., Drakoulis, D., Karetsos, G.: Optimazation of the Handover Algorithm based on the “Position of the Mobile Terminals”. In: Proceedings of Symposium on Communications and Vehicular Technology (October 2000)
8. Montavont, J., Noel, T.: IEEE 802.11 Handovers Assisted by GPS Information. IEEE, Los Alamitos (2006), 1-4244-0495-9/06
9. Huang, P.-J., Tseng, Y.-C.: A Fast Handoff Mechanism for IEEE 802.11 and IAPP Networks. In: IEEE 63rd Vehicular Technology Conference, VTC 2006-Spring. vol. 2, pp. 966–970 (2006)
10. Tseng, C.-C., Chi, K.-H., Hsieh, M.-D., Chang, H.-H.: Location-based Fast Handoff for 802.11 Networks. *IEEE Communication Letters* 9(4) (April 2005)
11. Huang, P.-J., Tseng, Y.-C., Tsai, K.-C.: A fast handoff mechanism for IEEE 802.11 and IAPP networks. In: IEEE 63rd Vehicular Technology Conference, VTC 2006 Spring, vol. 2, pp. 966–970 (2006)

Scalable Distributed Diagnosis Algorithm for Wireless Sensor Networks

Arunanshu Mahapatro and Pabitra Mohan Khilar

Department of CSE, National Institute of Technology, Rourkela, India
arun227@gmail.com, pmkhilar@nitrkl.ac.in

Abstract. This paper investigates the distributed self diagnosis problem for wireless sensor networks (WSN). One of the fundamental algorithm design issue for WSN is conservation of energy at each sensor node. A heartbeat comparison based diagnosis model is proposed, which is shown to be energy efficient. Analytical studies and simulation results show that the performance of proposed algorithm is comparable to that of the existing known algorithms in both delay and message count prospective. At the same time, the per-node message overhead is substantially reduced and becomes scalable.

Keywords: Distributed diagnosis, Comparison based model, WSN, scalable diagnosis.

1 Introduction

Over recent years, the market for WSN has enjoyed an unprecedented growth. WSNs are subject to tight communication, storage and computation constraint. All the nodes in WSN rely on batteries or other exhaustible means for their energy and wireless links continue to have significantly lower capacity. WSNs are expected to be deployed in inaccessible and hostile environments and thus more prone to failure. The availability of these sensor nodes, however, remains a major concern, if faulty sensor nodes are allowed to corrupt the network. System level diagnosis appears to be a viable solution to this problem.

Many authors have investigated this problem in their literature [1,2,3,4]. Article [1] presents a distributed fault detection algorithm for wireless sensor networks. The fault detection accuracy of a detection algorithm would decrease rapidly when the number of neighbor nodes to be diagnosed is small and the node's failure ratio is high. Article [4] address this problem by defining new detection criteria. Mourad Elhadeif *et al.* [5] proposed a distributed fault identification protocol (Dynamic-DSDP) for MANETs which assumes a static network. Dynamic-DSDP differs from Chessa and Santis model [6] in their dissemination strategies. It uses a spanning tree (ST) and a gossip style dissemination strategy, where the ST is created at each diagnosis period. Our protocol uses the same dissemination strategy, but avoids creation of ST during each diagnosis period. In this paper, a heartbeat based scalable diagnosis algorithm (SDDA) is proposed. The time and communication complexity is compared with [5,6].

2 System Model

The system under consideration accommodates n number of stationary homogeneous sensor nodes with unique identity number and same transmission range, which communicate via a packet radio network.

The proposed algorithm assumes: all nodes are fault free during deployment with a single sink node, static fault situation i.e no node is allowed to be faulty during algorithm execution. The communication algorithm ensures that: each sensor knows the identity of its neighbor, MAC protocol solves contention problem over logical link, the link level protocol provides one hop broadcast and one hop unicast routing, clock synchronization is achieved by periodical timing information exchanges through beacon frames. and communication channels between the nodes have bounded delay and flawless.

Communication Model: The communication graph of a WSN is represented by a digraph $G = (V, E)$, where V is set of sensor nodes in the network and E is set of edges connecting sensor nodes. Two nodes v_i and v_j are said to be adjacent only when the distance between them is less than the transmission range. For convenience the algorithm assumes that G is undirected that means $(v_i, v_j) \in E$ and $(v_j, v_i) \in E$. The send initiation time, T_{init} , is the time between a node initiating a communication and the last bit of the message being injected into the network. To simplify analysis, it is assumed that T_{send_init} is a constant. The minimum and maximum message delays, T_{min} and T_{max} , are the minimum and maximum times, respectively, between the last bit of a message being injected into the network and the message being completely delivered at a working neighboring node. [7]

Fault Model: Faults can be classified as either hard or soft fault. In hard-fault situation the sensor node is unable to communicate with the rest of the network, whereas a sensor with soft-fault continues to operate and communicate with altered behavior. This paper defines this altered behavior as random heartbeat sequence number, which does not match to heartbeat sequence number of other fault free sensors in WSN. This paper assumes only the permanent fault (hard and soft) situation, which uses a Spanning tree (ST) and a gossip style dissemination strategy. the spanning tree is created immediately after network deployment with sink node as root.

3 The Proposed Algorithm

In this section we introduce SDDA for WSN (see appendix), which is initiated by all the nodes simultaneously by sending a heartbeat message to its neighbors. A heartbeat message accommodates nodeID: the identification number of the node that initiated the heartbeat message and HB_seq_no: the physical sequence number of the heartbeat. In [6], nodes responds to each test request they receive, which increases the communication complexity and is suggested only when each

sensor is required to diagnose the fault status of its neighbors. However, the proposed algorithm responds only to the earliest arrived test request.

Complexity Analysis: Distributed diagnosis algorithms are usually evaluated with respect to their time complexity, space complexity and message complexity.

Theorem 1. *The message complexity is $O(n)$.*

Proof. The message cost associated with each message is as follows:

Message type	Message count	
Test	n	All nodes generates at most one test message.
Response	n	Each node responds to at most one test message.
Parent update	$n - 1$	If parent missing (faulty), it updates its parent field with a neighbor having lowest depth. In worst case $n - 1$ messages are exchanged.
Local dissemination	$n - 1$	Each node sends one message to its parent
Global dissemination	$n - 1$	each node sends one message to its child in global dissemination. Where in worst case the depth of the ST is $n - 1$.

Thus, the total message cost is $5n - 3$. □

Theorem 2. *The worst case time complexity is $(2n + 1)(T_{init} + T_{max}) + \psi$.*

Proof. Heartbeat generation phase ensures simultaneous initiation a heartbeat message simultaneously, which reaches at the neighboring nodes by at most $T_{init} + T_{max}$ time. In aggregation phase each node on reception of heartbeat message, evaluates the heartbeat sequence number and then initiates a response message. The farthest neighboring node receives this response message after $T_{init} + T_{max} + \psi$ amount of time, where ψ is the processing time and assumed constant. At the end of this phase, nodes with faulty parents send the adopt request which needs at most $T_{init} + T_{send_max} + \psi$ time. In local disseminating phase each node sends its own diagnostic to its parent. The parent collects all diagnostics of its children and merge these diagnostics to its own diagnostic. In worst case the depth of ST is $n - 1$, hence the worst case time complexity of this stage is $(n - 1)(T_{init} + T_{max}) + \psi$. In global disseminating phase the root node disseminates the global view that reaches at the leaf node with highest depth costing time $(n - 1)(T_{init} + T_{max})$. Thus, the total time is $(2n + 1)(T_{init} + T_{max}) + \psi$. □

Simulation Results and Comparison With Related Works

The proposed algorithm was simulated on randomly generated network of size n . A set of communication graph G were generated with a known connectivity k . T_{init} , T_{min} and T_{max} were kept fixed at $10\mu s$, $20\mu s$ and $25 ms$. Simulation time is set to 200 seconds. Simulations were done using discrete event simulation techniques, where nodes were initially given one unit of energy.

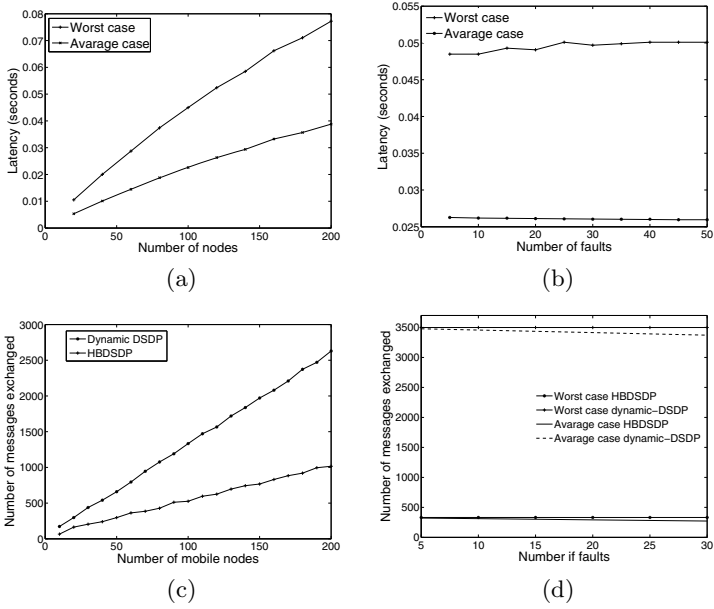


Fig. 1. (a) Latency in diagnosing failure events. (b) Time complexity of HBDSDP in presence of faults. (c) Message count in diagnosing failure events. (d) Message complexity of HBDSDP in presence of faults.

Table 1. Comparison with related work

	Message complexity	Time complexity
Chessa <i>et al.</i> model	$nd_{max} + n(n + 1)$	$\delta_G T_{gen} + \delta_G T_f + T_{out}$
Dynamic-DSDP	$nk + 3n - 1$	$\delta_G T_{gen} + 3d_{ST} T_f + 2T_{out}$
HBDSDP	$5n - 3$	$(2n + 1)(T_{init} + T_{max}) + \psi.$

d_{max} : The maximum of the node degree δ_G : The diameter of graph G.

T_f : The upper bound to the time needed to propagate a dissemination message

T_{gen} : An upper bound to the elapsed time between the reception of the first diagnostic message and the generation of the test request. d_{ST} : Depth of the spanning tree.

4 Discussion

This paper addresses the fundamental problem of identifying faulty (soft and hard) sensors in WSN. Both analytical and simulation study of the proposed algorithm is presented. The use of heartbeat based approach, further reduces the number of bits exchanged per message. Both the message and time complexity of the algorithm is $O(n)$ for an n-node WSN. An interesting open question is whether a self diagnosis algorithm for dynamic fault situation with lower message cost can be developed that can either have same or less latency. In the future work we are investigating this open question.

References

1. Lee, M.-H., Choi, Y.-H.: Fault detection of wireless sensor networks. *Computer Communications* 31(14), 3469–3475 (2008)
2. Krishnamachari, B., Iyengar, S.: Distributed bayesian algorithms for fault-tolerant event region detection in wireless sensor networks. *IEEE Transactions on Computers* 53(3), 241–250 (2004)
3. Luo, X., Dong, M., Huang, Y.: On distributed fault-tolerant detection in wireless sensor networks. *IEEE Transactions on Computers* 55(1), 58–70 (2006)
4. Jiang, P.: A new method for node fault detection in wireless sensor networks. *Sensors* 9(2), 1282–1294 (2009)
5. Elhadef, M., Boukerche, A., Elkadiki, H.: A distributed fault identification protocol for wireless and mobile ad hoc networks. *Journal of Parallel and Distributed Computing* 68(3), 321–335 (2008)
6. Chessa, S., Santi, P.: Comparison-based system-level fault diagnosis in ad hoc networks. In: *Proceedings of 20th IEEE Symposium on Reliable Distributed Systems*, pp. 257–266 (2001)
7. Subbiah, A., Blough, D.M.: Distributed diagnosis in dynamic fault environments. *IEEE Transactions on Parallel and Distributed Systems* 15(5), 453–467 (2004)

Appendix

The proposed scheme comprises of three main stages like heartbeat generation phase, aggregation phase, disseminating phase.

Algorithm 1. Heartbeat Generation Phase

- 1: {HBrequest is initialized to FALSE and Set to TRUE once the sensor u generates its HBrequest.}
 - 2: **if** $HBrequest == FALSE$ **then**
 - 3: $HB_seq_no = 1$;
 - 4: Broadcast($nodeID, HB_seq_no$); $HB_seq_no ++$; $HBrequest = TRUE$;
 - 5: **else**
 - 6: **if** $HB_seq_no == Max_seq_no$ **then**
 - 7: $HB_seq_no = 1$;
 - 8: **end if**
 - 9: Broadcast($nodeID, HB_seq_no$); $HB_seq_no ++$;
 - 10: **end if**
 - 11: SetTimer(T_{out});
-

Algorithm 2. Aggregation Phase

```

1: {Message has been sent by sensor  $v \in N(w)$ . ResponseFlag is initialized to FALSE}
2: repeat
3:   if ( $v.HB\_seq\_no \neq w.HB\_seq\_no - 1$ ) OR ( $v.HB\_seq\_no \neq w.HB\_seq\_no$ ) then
4:      $F_w = F_w \cup \{v\}$ ; {Message from a faulty node: may be a soft fault}
5:   else
6:     if ( $v.HB\_seq\_no == w.HB\_seq\_no - 1$ ) AND ( $v \notin F_w$ ) AND  $Flag == FALSE$  then
7:       Increment and broadcast  $HB\_seq\_no$  and  $Flag = TRUE$ ;
8:     end if
9:     if  $v.HB\_seq\_no == w.HB\_seq\_no$  then
10:       $FF_w = FF_w \cup \{v\}$ ;  $Node_w[v].status = working$ ;
11:    else
12:       $F_w = F_w \cup \{v\}$ ;
13:    end if
14:    if  $T_{out} == TRUE$  then
15:       $F_w = F_w \cup \{N(w) - (F_w \cup FF_w)\}$ ;
16:    end if
17:  end if
18: until ( $F_w \cup FF_w \neq N(w)$ )
19: if  $w.parent \in F_w$  then
20:   {Find the node with lowest depth from  $FF_w$  and declare it as new parent of  $w$ }
21: end if

```

Algorithm 3. Disseminating Phase

```

1: {LocalDiagnosis and GlobalDiagnosis is initialised to FALSE.}
2: repeat
3:   if  $w.children == NULL$  then
4:     Unicast( $parent, F_w, Node_w$ )
5:   end if
6:   if  $v \in w.children$  then
7:      $Node_w = Node_w \cup Node_v$ ;  $F_w = F_w \cup F_v$ ;  $Children = Children \cup \{v\}$ ;
8:     if  $w.children == Children$  then
9:       Unicast( $parent, F_w, Node_w$ );
10:    end if
11:  end if
12:  if  $w == initiator$  then
13:    Broadcast( $F_w, Node_w$ ); LocalDiagnosis = TRUE;
14:  end if
15: until (LocalDiagnosis == FALSE)
16: repeat
17:   if  $w.children \neq NULL$  then
18:      $Node_w = Node_w \cup Node_v$ ;  $F_w = F_w \cup F_v$ ; Broadcast( $F_w, Node_w$ );
19:   end if
20:   if  $w.Depth == ST\_Depth$  then
21:     GlobalDiagnosis = TRUE;
22:   end if
23: until (GlobalDiagnosis == FALSE)

```

Comparison of Routing Protocols for MANET and Performance Analysis of DSR Protocol

Parma Nand* and S.C. Sharma

Wireless Computing Research Lab, Indian Institute of Technology, Roorkee
astya2005@gmail.com, {astyadpt,scs60fpt}@iitr.ernet.in

Abstract. MANET, a multihop mobile adhoc network is comprised of mobile nodes which communicate over radio. Each host is equipped with a CSMA/CA (carrier sense multiple access with collision avoidance) transceiver. Mobile ad-hoc networks are suitable for temporary communication links as they do not require fixed infrastructure. The communication among routers is difficult due to its frequent changing network topology and requires efficient and dynamic routing protocol. This presents comparison of popular broadcast based proactive, reactive and hybrid adhoc routing protocols. Further, the performance of Dynamic Source Routing (DSR), an on-demand, routing protocol based on IEEE 802.11 is examined on performance metrics throughput, end-to-end packet delay, jitter and salvaged packets using QualNet 5.0.2 network simulator.

Keywords: Adhoc networks, wireless networks, MANET, CBR, route discovery, simulation, performance evaluation, MAC, IEEE 802.11, DSR.

1 Introduction

A mobile multi-hop ad-hoc network (MANET) is a set of mobile nodes which communicate over radio and do not need any infrastructure. These networks are very flexible and suitable for several types of applications, as they allow the establishment of temporary communication without any pre installed infrastructure. Due to the limited transmission range of wireless interfaces, in most cases communication has to be relayed over intermediate nodes. Thus, in mobile multi-hop ad-hoc networks each node also has to be a router [1]. Beside the disaster and military application domain the deployment of mobile ad-hoc networks for multimedia applications is another interesting domain. To find a route between the communication end-points is a major problem in mobile multi hop ad-hoc networks. Many different approaches to handle this problem are proposed in recent years, but so far no routing algorithm has been found suitable for all situations.

In this paper the comparison of popular proactive, reactive and hybrid adhoc routing protocols is presented. In literature some of the performance analysis and comparisons [2,3,4] are reported using ns2. Further, the performance analysis of DSR routing protocol based on IEEE 802.11[5] is analyzed. This paper explores the impact of MAC overhead and multiple hops on achievable data throughput, jitter, end-to-end delay and salvaging with varying time and mobility using Qualnet simulator [6].

* Corresponding author.

2 Routing Protocols in MANETs

Routing is the process of finding a path from a source to some arbitrary destination on the network. The broadcasting [7,8,9] is inevitable and a common operation in ad-hoc network. It consists of diffusing a message from a source node to all the nodes in the network. The routing protocols are classified as follows.

Proactive protocols, also called table driven, continuously evaluate and maintain consistent and up to date routing information within the network, so that when a packet needs to be forwarded the route is already known and can be used immediately. The table driven protocols for example are

1. Destination sequenced Distance vector routing (DSDV)[10]
2. Optimized Link State Routing (OLSR) [11]

Reactive routing protocols, also called on demand, invoke a route determination procedure only on demand. A node wishing to communicate with another node first seeks for a route in its routing table. If it finds one, the communication starts immediately, otherwise the node initiates a route discovery. For example

1. Ad-Hoc On-demand Distance Vector (AODV) [12].
2. Dynamic Source Routing (DSR) [13,14]

Hybrid protocols combine the better features of both proactive and reactive routing protocol. For example

1. Temporally ordered routing algorithm(TORA)[15]
2. Zone Routing Protocol (ZRP)[16]

Characteristic summery of popular protocols of these categories is given in Table 1.

Table 1. Characteristic summery of DSDV, DSR, TORA Protocols

Protocol	Destination- Sequenced Distance- Vector (DSDV) [10]	Dynamic Source Routing (DSR) [13,14]	Temporally ordered routing algorithm (TORA) [14]
Category	Proactive	Reactive	Hybrid
Metrics	Shortest path	Shortest path, next available	Shortest path, next available
Route Recovery	Periodic broadcast	New route, notify source	Reverse link
Route repository	Routing table	Route cache	Routing table
Broadcasting	Simple	Simple	Simple
Loop Free	Yes	No provision for it	Yes
Communication Overhead	High	High	High
Feature	Distributed algorithm	Completely on demand	Control packets localized to area of topology change

3 Dynamic Source Routing Protocol

DSR [13,14] is a source routing, on-demand, routing protocol. The source (sender) knows the complete hop-by-hop route to the destination. These routes are stored in a route cache. The data packets carry the source route in the packet header. It is composed of two parts Route Discovery and Route Maintenance.

3.1 Route Discovery

When a node in the adhoc network attempts to send a data packet to a destination for which route is not known, it uses a route discovery process to find a route. Route discovery uses simple flooding technique in the network with route request (RREQ) packets. Each node receiving an RREQ rebroadcasts it further, unless it is the destination or it has a route to the destination in its route cache. Such a node replies to the RREQ with a route reply (RREP) packet that is routed back to the original source. RREQ and RREP packets are also source routed. RREQ builds up the path traversed so far. The RREP routes itself back to the source by traversing this path backward, the route carried back by the RREP packet is cached at the source for future use.

3.2 Route Maintenance

The periodic routing updates are sent to all the nodes. If any link on a source route is broken, the source node is notified using a route error (RERR) packet. The source removes any route using this link from its cache. A new route discovery process must be initiated by the source if this route is still needed. Following techniques are used to clean up the caches of other nodes to improve performance.

- (i) Salvaging: An intermediate node can use an alternate route from its own cache, when a data packet meets a failed link on its source route.
- (ii) Gratuitous route repair: A source node receiving RERR piggybacks the RERR in the following RREQ, to clean the caches of other nodes that may use failed link.

4 Simulation Setup

The Qualnet 5.0 simulator is used to analyze DSR protocol. In analysis UDP (User Datagram Protocol) connection is used and over it CBR (Constant bit rate) is applied between source and destination. The 54 nodes are placed uniformly initially. The random waypoint mobility model with the maximum speed of 50 mps is used in a rectangular field. The multiple CBR application are applied over 11 different source nodes—6,1,34,3,33,17,7,38,39,52,30 and destinations nodes—31,36,4,25,18,37, 44,43, 54,42, 45 respectively. The simulations parameters are shown in table 2.

4.1 Performance Metrics

Throughput: Throughput is the average rate of successful data packets received at destination. It is usually measured in bits per second (bit/s or bps), and sometimes in data packets per second.

End-to-End Delay: It is the accumulation of processing and queuing delays in routers, propagation delays, and end-system processing delays. The packets that are delayed by more than the threshold value are effectively lost.

Jitter: Jitter is the variation of the packet arrival time. In jitter calculation the variation in the packet arrival time is expected to be low. The delays between the different packets need to be less than threshold value for better performance.

Salvaging: An intermediated node uses an alternative route from its cache, when a data packet meets a failed link on its source route. This saves the packets and improves the routing performance.

Table 2. Simulation Parameters

Parameter	Value
Area	1500mX1500m
Simulation Time	90,120, 200 sec
Bandwidth	22.0e6
Data rate	2.0e06
Path Loss Model	Two Ray Model
Packet size	1024 bytes
Mobility Model	Random-Way Point
Node Speed	Maximum 50 mps
Physical Layer Radio type	IEEE 802.11b
MAC Protocol	IEEE 802.11
Antenna Model	Omni-directional

5 Results and Discussion

The performance of DSR is analyzed with varying mobility speed, traffic load and simulation times using Qualnet 5.0.2. The snapshot of broadcasting, nodes mobility and transmission of data is shown in figure 1. The simulation results are shown in figures from 2 to 5.

Salvaging: Performance is analyzed on this metric as the DSR routing protocol is having this very useful parameter to improve performance by choosing alternative routes from cache in case of link breakage. Simulation performance shows that the good number of packets is saved by the salvage technique of DSR protocol and this help in improving the routing performance. It is observed that best salvaged are found for 120 sec of simulation time because of limited node mobility as shown in figure 2.

Throughput: In simulation successful packet delivery is observed with increasing MAC based CBR data rate and varying mobility and simulation time. The throughput is found to vary at nodes because of the mobility of nodes that leads to link failures. Average throughput is found to be better during 120 sec of simulation time with average of mobility and quick reuse of links from the cache in comparison to other simulation timings. It is shown in figure 3.

End-to-End Delay: The mobility causes nodes reachability problem and hence aggravates propagation and processing delays. The high end-to-end delay of DSR is also attributed to the failures of links due to high mobility and lack of explicit mechanism for expire stale routes in the cache or prefer fresher routes. This is reflected in figure 4 as it is more for the 200 sec of simulation time than that for simulation time of 90 sec.

Jitter: The simulation analysis shows that the delay in packet arrival more in general and especially during 200 sec simulation time. The jitter is found to be very high at the node 31 because of its mobility during 120 sec of simulation time as shown in figure 5. The high jitter is attributed to the aggressive and excessive use of cache by the DSR protocol and inability to prefer fresher routes.

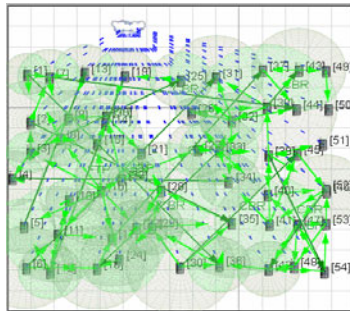


Fig. 1. Animation view of simulation

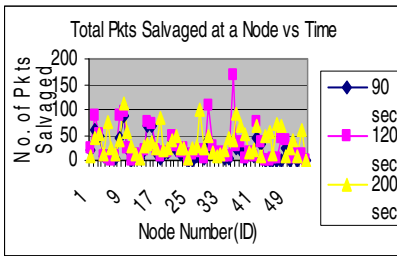


Fig. 2. Total Packets Salvaged at a Node vs Time

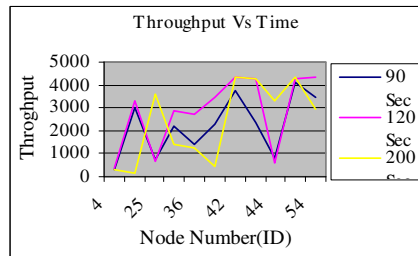


Fig. 3. Throughput vs Time

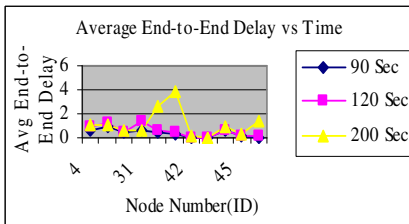


Fig. 4. Average End-to-End Delay vs Time

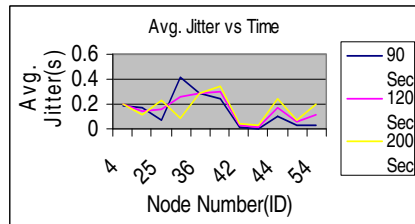


Fig. 5. Average Jitter vs Time

6 Conclusion

The DSR performs better for the scenario having low mobility and performs poor otherwise because of its aggressive use of cache. The caching of alternative route per destination is also one of the factors responsible for its average performance in case of high node mobility. The poor performances of DSR is mainly attributed to the aggressive use of caching, and lack of proper mechanism to expire stale routes or to determine the freshness of routes when multiple choices are available and therefore the jitter and the average end-to-end delay are also high. It is found that the throughput is average with the mobility of nodes because of the reasons mentioned above. The salvaging technique of DSR protocols is very helpful in saving the packets and hence improving its performance due to its cache strategy.

Therefore, with simulation analysis it is observed that DSR performs better with low mobility, lower routing overhead and in less stressful situation.

References

1. Toh, C.-K.: *Adhoc Mobile Wireless Networks: Protocols and Systems*. Prentice Hall, Englewood Cliffs (2002)
2. Yadav, N.S., Yadav, R.P.: Performance Comparison and Analysis of Table Driven & On Demand Routing Protocols for Mobile Adhoc Networks. *International Journal of Information Technology* 4(2), 101–109 (2007)
3. Pirzada, A.A., McDonald, C., Datta, A.: Performance Comparison of Trust-Based Reactive Routing Protocols. *IEEE Transactions on Mobile Computing* 5(6), 695–710 (2006)
4. Belding-Royer, E.: Royer, Routing approaches in mobile ad hoc networks. In: Basagni, S., Conti, M., Giordano, S. (eds.) *Ad Hoc Networking*, IEEE Press, Wiley (2003)
5. IEEE, 1997, *Wireless LAN Medium Access Control (MAC) and Physical layer PHY Specifications*, IEEE Std. 802.11 (1997)
6. Qualnet Simulator, <http://www.scalable-networks.com>
7. Ni, S.Y., Tseng, Y.C., Chen, Y.S., Sheu, J.P.: The broadcast storm problem in a mobile ad hoc network. In: *Proceedings of the 1999 Fifth Annual ACM/IEEE International Conference on Mobile Computing & Networking*, August 1999, pp. 151–162. IEEE Computer Society, New York (1999)
8. Zhang, Q., Agrawal, D.P.: Dynamic probabilistic broadcasting in MANETs. *Journal of Parallel and Distributed Computing* 65(2), 220–233 (2005)
9. Williams, B., Camp, T.: Comparison of broadcasting techniques for mobile ad hoc networks. In: *Proceedings of the ACM International Symposium on Mobile Ad Hoc Networking and Computing (MOBIHOC 2002)*, pp. 194–205 (2002)
10. Perkins, C., Bhagwat, P.: Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) Routing. In: *SIGCOMM 1994 Comp. Comm. Review*, vol. 24(4), pp. 234–244 (October 1994)
11. Clausen, T., Jacquet, P., Laouiti, A., et al.: *Optimized Link State Routing Protocol*, draft-ietf-manet-olsr-06.txt (September 2002)
12. Perkins, C., Royer, E., Das, S.: *Adhoc on demand distance vector (AODV) routing*, IETF RFC No. 3561 (July 2003)

13. Broch, J., Johnson, D., Maltz, D.: The dynamic source routing protocol for mobile adhoc networks for IPv4 IETF RFC 4728 (February 2007)
14. Johnson, D., Maltz, D.: Dynamic source routing in adhoc wireless networks. In: Imielinski, T., Korth, H. (eds.) Mobile computing, ch. 5. Kluwer Academic, Dordrecht (1996)
15. Park, V., Corson, S.: Temporally-Ordered Routing Algorithm (TORA) ver 1 Functional Specification, draft-ietf-manet-tora-spec-04.txt (July 2001)
16. Haas, Z.J., Pearlman, M.R., Samar, P.: The Zone Routing Protocol (ZRP) for Ad Hoc Networks, draft-ietf-manet-zone-zrp-04.txt (July 2002)

Transmitter Based Capacity Enhancement with Cross-Layer Design Approach for IEEE 802.11 Ad-hoc Networks

Satish Ket and R.N. Awale

Veermata Jijabai Technological Institute, Mumbai, India
satishket@gmail.com, rnawale@vjti.org.in

Abstract. We propose the Transmitter Based Capacity Enhancement Algorithm for Wireless Ad-hoc Networks using Cross-Layer Design Approach (TCECLD) by dynamically adapting the data rate. The proposed algorithm uses the attributes of Auto-Rate Fallback (ARF) [1] and Receiver Based Auto-Rate (RBAR) [2] control mechanisms with additional practical features to facilitate multipath fading. Like ARF, Acknowledgement (ACK) feedback count is used along with received signal strength for dynamic rate selection. Unlike RBAR, no modification of the Medium Access Control (MAC) frame format is necessary for feedback control since the channel state information is conveyed through Cross-Layer interface. The mechanisms are described to implement the proposal on top of the existing Auto-Rate adaptation scheme in a nearly IEEE 802.11 compliant manner. We also analytically study and characterize the gains in throughput as a function of the channel conditions. Finally an extensive set of simulations are performed on IEEE 802.11b media access protocols to observe the system throughput.

Keywords: TCECLD, ARF, Wireless Ad-hoc Networks.

1 Introduction

Wireless local area networks are becoming increasingly popular. This is due to the ratification of standards like IEEE 802.11 [3], which have laid the foundation for off the shelf wireless devices capable of transmitting at high data rates. Higher data rates are commonly achieved by more efficient modulation schemes. The IEEE 802.11a and 802.11b media access protocols provide a physical layer multi-rate capability with the set of possible data rates as 6, 9, 12, 18... 54 Mbps and 1, 2, 5.5 & 11 Mbps respectively.

As the multi-rate IEEE 802.11 enhancements are physical layer protocols, MAC mechanisms are required to exploit this capability. The ARF protocol [1] was the first commercial implementation of a MAC that utilizes this feature. Since the 802.11 standard does not specify any algorithm and/or protocol to efficiently utilize the multiple transmission rates, many rate adaptation schemes have been proposed [1, 2], [5-10]. The effectiveness of a rate adaptation scheme depends not only on how fast it can respond to the variation of wireless channel but also on how the collisions may be

detected and handled in a multi-user environment where frame collisions are inevitable due to the contention nature of the 802.11 DCF (Distributed Coordination Function). In addition it must differentiate between frame collision and channel error as discussed in [10].

In this paper, an enhanced protocol for multi-rate IEEE 802.11 in Wireless Ad-hoc Networks is proposed. The key idea is to exploit high quality channels when they occur, via transmission with higher rates. The channel access is granted along with the selection of data rate as per the received signal strength at that moment. Consequently, nodes transmit at higher rates under high quality channels than under low quality channels. Two mechanisms are required to realize the proposal. First, it requires a multi-rate MAC protocol such as ARF to access the medium at rates above the base rate which can be applied to both transmitter and receiver based protocols. Second, it requires a mechanism to know the channel condition by the recently received signal strength which is made available with Cross-Layer Communication [11,12]. As the transmitter decides the data rate, the proposal is named as TCECLD.

To study the performance of the proposal, an analytical model is adopted that characterizes the throughput gains as compared to IEEE 802.11 as a function of the physical layer channel conditions. Finally, the extensive simulation study is performed to evaluate the proposal in realistic scenarios and to isolate the performance factors that determine throughput gains. Only ad-hoc network scenarios are considered. Example findings are as follows. (1) In most cases, the throughput gain is about 2 to 2.5 times more compared to ARF. (2) The throughput gain of TCECLD is about 1.5 times more as compared to Collision Aware Rate Adaptation (CARA) [10]. (3) Ultimately load carrying capacity of the Ad-hoc network increases by TCECLD.

The rest of the paper is organized as follows. Related work is presented in Section 2. Channel model and adopted Analytical model are discussed in Sections 3 and 4 respectively. Section 5 presents the proposal. Simulation results and performance analysis are discussed in Section 6, and finally the paper concludes in Section 7.

2 Related Work

There have been remarkable studies on the capacity improvement by rate adaptation in the 802.11 WLANs. A transmitter station can change its transmission rate with or without feedback from the receiver, where the feedback information could be either Signal-to-Interference/Noise Ratio (SINR) or the desired transmission rate determined by the receiver.

In [2], after the receiver specifies its desired transmission rate and feeds back to the transmitter as part of a modified RTS (Ready to Send)/CTS (Clear to Send) exchange; the transmitter adapts its transmission rate accordingly. The rate adaptation is dictated by the receiver in this approach. In few of the approaches [1,10], a transmitter station makes the rate adaptation decision solely based on its local ACK information. In the 802.11 standard, an ACK frame is transmitted by the receiver upon successful reception of a data frame. The transmitter assumes a successful delivery of the

corresponding data frame only after receiving an ACK frame correctly. On the other hand, if an ACK frame is received in error or no ACK frame is received at all, the transmitter assumes failure of the corresponding data frame transmission. In [7] the high quality channels are exploited by sending multiple back-to-back packets. In [9,10] the data rate is decided based on local channel estimation made during ACK frame receptions, assuming a symmetric wireless channel between the transmitter and the receiver. In such cases a very good performance is observed, but usually require extra implementation efforts. In [1,5,9], the local ACK information is used while selecting the transmission rate, which is very simple to implement. It has been pointed out in [9] that the fundamental issue while designing a rate adaptation scheme is the timings of when to increase and when to decrease the transmission rate. The effectiveness of a rate adaptation scheme depends greatly on how fast it may respond to the wireless channel variation. The schemes presented in [5,9] addresses this issue and enhance the original ARF by allowing a transmitter station to increase its rate in an adaptive manner over a time varying wireless channel.

3 Channel Model

The transmitted radio frequency signal is reflected by both natural and man made objects. Thus, the signal at the receiver is a superposition of different reflections of the same signal received with varying delays and attenuations. Based on the relative phases of different reflections at the receiver, the different copies of the same signal may add coherently or tend to cancel out. Coherent addition of the copies can result in large received signal powers and cancellation eventually leads to zero received signal power. An accurate and widely utilized model which considers time varying multi-path propagation [4] is

$$y(t) = \sum_{i=1}^{p(t)} A_i(t) x(t - \tau_i(t)) + z(t) \quad (1)$$

where $x(t)$ is the transmitted signal and $y(t)$ is the received signal. The time-varying multi-path propagation is captured by the attenuation of each path $A_i(t)$, the time delays $\tau_i(t)$ and the number of paths $p(t)$. The additive term $z(t)$ is generally labeled as the background noise and represents the thermal noise of the receiver. Note that the loss suffered by the signal during its propagation along different paths is captured in $A_i(t)$, and depends on the distance between the sender and the receiver.

Recognizing that the received SNR (Signal to Noise Ratio) can be used to capture the packet level performance of any physical layer implementation. The following model is used for the received SNR for transmitter power P at packet transmission time t_p ,

$$SNR(t_p) = Pd(t_p) - \beta \frac{\rho(t_p)}{\sigma^2} \quad (2)$$

where $d(t_p)$ is the distance between the sender and the receiver at time t_p , β is the path loss exponent, $\rho(t_p)$ is the average channel gain for the packet at time t_p , and σ variance of the background noise $z(t)$.

The short time-scale variation in the received SNR is captured by the time-varying parameter $\rho(t_p)$, known as the fast fading component of the fading process. The time-variation of $\rho(t_p)$ is typically modeled by a probability distribution and its rate of change [4]. An accurate and commonly used distribution for $\rho(\cdot)$ is the Ricean distribution,

$$p(\rho) = \frac{\rho}{\sigma^2} e^{-\left(\frac{\rho}{2\sigma^2} + K\right)} I_0(2K\rho) \tag{3}$$

where K is the distribution parameter representing the strength of the line of sight component of the received signal and $I_0(\cdot)$ is the modified Bessel function of the first kind and zero-order [4]. For $K = 0$, the Ricean distribution reduces to the Rayleigh distribution, in which there is no line-of-sight component.

4 Analytical Model

In this paper for the analytical evaluation of the saturation throughput, the same method as suggested by Binachi [13] is used, assuming the ideal channel conditions with no hidden terminals. In the analysis, it is assumed that a fixed number of stations, each always having a packet available for transmission. In other words, operation is in *saturation* conditions.

The normalized system throughput S is represented as;

$$S = \frac{E[\text{payload_information_transmitted_in_a_slot_time}]}{E[\text{length_of_a_slot_time}]}$$

Let $E[P]$ be the average packet payload size, the average amount of payload information successfully transmitted in a slot time is $P_{tr}P_sE[P]$, since a successful transmission occurs in a slot time with probability $P_{tr}P_s$ and with probability $P_{tr}(1-P_s)$ it contains a collision. Hence S becomes;

$$S = \frac{P_{tr}P_sE[P]}{(1-P_{tr})\sigma + P_{tr}P_sT_s + P_{tr}(1-P_s)T_c} \tag{4}$$

where,

$$P_{tr} = 1 - (1 - \tau)^n,$$

$$P_s = \frac{n\tau(1 - \tau)^{n-1}}{P_{tr}},$$

$$T_s = RTS + CTS + H + E(P) + ACK + DIFS + 3SIFS + 3\delta,$$

$$T_c = RTS + SIFS + \delta.$$

The probability τ that a station transmits in a randomly chosen slot time can be found by solving the two equations;

$$\tau = \frac{2(1-2p)}{(1-2p)(W+1) + pW(1-(2p)^m)} \text{ and}$$

$$p = 1 - (1-\tau)^{n-1}$$

where $W = CW_{min}$, $CW_{max} = 2^m CW_{min}$, p is the probability that each packet collides at each transmission attempt, and regardless of the number of retransmission attempts, m is the back-off stage and n is the total contending nodes.

Assumed that each packet is transmitted by means of the RTS/CTS Access mechanism and it is obvious that in such a case, collision can occur only on RTS frames. Also as this RTS and CTS are exchanged with the rate 2Mbps, T_s , T_c , $E[P]$ and σ are constants. Hence the throughput expression depends on τ , in turn the network size n . The value of τ is approximated as;

$$\tau = \frac{1}{n\sqrt{\frac{T_c}{2\sigma}}} \tag{5}$$

The value of τ is solely dependent on the number contending nodes in the network.

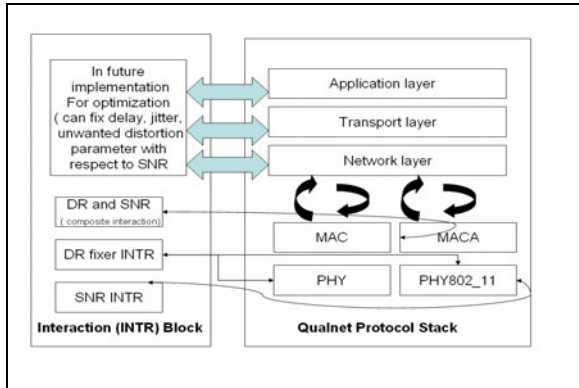


Fig. 1. Model Implementation in MAC with Cross-Layer Approach

5 The Proposal

A generic model for cross-layer interactions is proposed [Fig. 1]. Interaction block keeps a track of the instantaneous value of SNR, which is grabbed from received signal in the Physical layer. This value is accessible by all the layers as and when required for specific decision making. The present work is restricted up to MAC layer only. However, this can be extended to upper layers. In 802.11 the data rate of transmission is decided at MAC layer and instantaneous SNR value reflects the channel condition in time varying and mobile environment. Exploitation of channel

condition to increase the network capacity is the aim, which is achieved by this additional Cross-Layer interaction without changing the basic architecture of the Protocol Stack as seen in [12]. The optimization is achieved in the system by adjusting the data rate using threshold-based technique [7]. In a threshold-based scheme, the rate is chosen by comparing the received SNR value of the signal against an array of thresholds representing performance limits of the available modulation techniques. The modulation technique with the highest data rate for the estimated SNR value is chosen.

The selected modulation technique results in the feasible data rate to be used in subsequent transmissions. Let $p_1, p_2, p_3, \dots, p_{m-1}$ are SNR thresholds for different suitable rate limits. For example, p_1 indicates that if the received SNR level is below p_1 , rate r_1 is feasible. In case the received SNR level is above p_1 but below p_2 , rate r_2 is feasible and so on. A region surrounded by two subsequent SNR thresholds which is suitable for a particular rate.

```

m = 0, n = 0; //Initialization.
while (transmission queue is nonempty) {
    RTS/CTS Exchange;
    if (is Packet Transmitted?) then successful ();
    else failure ();
}

successful () {
    increment m; reset n;
    if (m >= mthr) then {
        data_rate = ri; // As per SNRi value.
        reset m;
    }
}

failure () {
    increment n; reset m;
    if (n >= nthr) then {
        data_rate = ri--; // Next lower value from data rate set.
        reset n;
    }
}

```

Fig. 2. TCECLD Algorithm

5.1 Transmitter Based CECLD Algorithm

In a fully connected ad-hoc topology in which all nodes are in radio range of each other, base rate IEEE 802.11 indeed provides long-term fairness. If multi-rate is adopted, still identical long-term time shares can be obtained but at different throughputs. For example, suppose there are two flows, one with a low signal strength such that it can only transmit at the base rate of 2 Mbps and the other

with a high signal strength so that it can transmit at rate 11Mbps. Thus, in contrast to the focus on throughput fairness of which attempt to normalize flow throughputs, temporal fairness is more suitable for multi-rate networks as normalizing flow throughputs would cancel the throughput gains available due to a multi-rate physical layer. To improve the system performance in terms of throughput, i.e. to improve the system capacity TCECLD algorithm for wireless ad-hoc networks is proposed.

In TCECLD the parameter tuned is data rate. As shown in Fig. 1, additional interface is created between Physical and MAC layer. Channel condition is estimated by recent SNR value of received signal. The algorithm for data rate selection for next attempt of transmission/retransmission is given in Fig. 2. Table 1 gives the list of notations used in the algorithm.

Table 1. List of Notations Used in the Algorithm

Notations	Comments
m	Consecutive success count
n	Consecutive failure count
m_{thr}	Consecutive success threshold
n_{thr}	Consecutive failure threshold
r_i	Data rate i , from 802.11b data rate set, {1,2,5.5,11}
SNR_i	SNR value at instant i

Table 2. System Parameters

Parameter	Value
Packet Payload	4096 bits
MAC Header	272 bits
PHY Header	288 bits
RTS	160 bits + PHY Header
CTS	112 bits + PHY Header
ACK	112 bits + PHY Header
Packet Arrival Rate	2/3 Mbps
Propagation Delay	1 μ s
Slot Time	20 μ s
SIFS	10 μ s
DIFS	50 μ s

6 Performance Evaluation

In this section, the effectiveness of the proposal is evaluated by using the Qualnet 4.5 after enhancing the original 802.11 DCF module to support the 802.11b Physical and the time varying wireless channel model.

Mainly 802.11b ad-hoc networks are simulated. Equations (1) and (2) are used for the channel and SNR estimations respectively. Each station transmits with 15 dBm power, and all the stations are static unless stated within the range of each other. Two Ray Path-loss model [4] is used to simulate the environments. Moreover, the multi-path fading effect is considered with which the channel condition between the transmitter and receiver varies over the time. The Ricean fading model [14] is used to simulate the time varying wireless channel conditions. The Ricean distribution is given by Equation (3) as addressed in Section 3. The consecutive success threshold (m_{thr}) is set to 2, and the consecutive failure threshold (n_{thr}) to 2, for CARA and TCECLD. Simulations under various network topologies and network size are conducted. Each node transmits in a greedy mode, i.e. its data queue is never empty and all the data frames are transmitted without fragmentation. The data payload length is 512 bytes unless specified otherwise. The system parameters are given in Table 2.

6.1 Saturated Throughput for 1 to 1 Node Topology

TCECLD is tested with the simplest 1 to 1 node topology in which nodes continuously transmit packets to each other. The offered load of the system is increased up to 6 Mbps. From Fig. 3 it is observed that the performance of CARA and TCECLD is almost similar. In this scenario the nodes are stationary and no time varying channel is used. CARA and TCECLD adapt different mechanisms to select the data rate but because of good channel conditions both perform similar and achieve the maximum data rate within a very short time and exploit the channel up to its maximum capacity. Due to limitations of ARF, it performs poorly. The saturation throughput achieved by CARA and TCECLD is almost doubled than ARF.

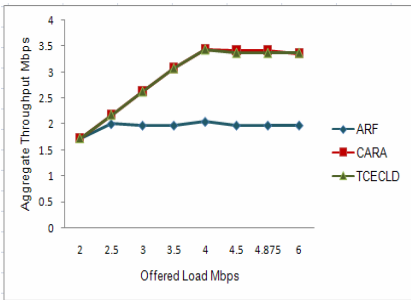


Fig. 3. Aggregate Throughput as a function of Offered Load

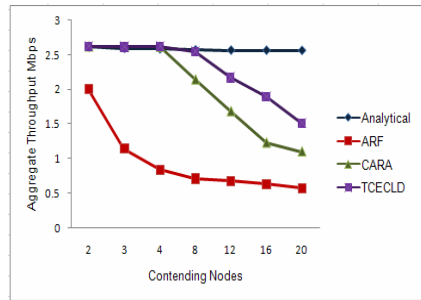


Fig. 4. Aggregate Throughput as a function of Contending Nodes in Stationary Channel Condition

6.2 Fully Connected Topologies with Varying Number of Contending Stations

Fully connected ad-hoc networks with varying number of contending nodes are considered in order to study the system performance. In this scenario, various number of contending nodes are evenly spaced in a terrain of 500mX500m making sure that all are within each others range and there are no hidden nodes and all are static. The wireless channel model is not time varying. The nodes are transmitting the data packets at 3 Mbps. Simulation results are plotted in Fig. 4. TCECLD gives the throughput similar to Analytical throughput calculated by Equations (4) and (5) up to the network size of 8 contending nodes. Whereas in case of CARA it is restricted up 4 nodes. In case of ARF, the aggregate system throughput is degraded severely even with a small number of contending nodes in the network. For example, when the number of contending nodes increases from 2 to 4, the aggregate throughput with ARF drops from 2 Mbps to about 0.8 Mbps and for further increase in contending nodes, it drops to about 0.5 Mbps. CARA does not work as poorly as ARF. However, it gives inferior performance than TCECLD since it changes the data rate gradually.

There are two main reasons for the poor performance of ARF. First, since ARF waits for 10 successful transmissions to increase the rate by one step, even channel condition is far better and a wireless station may decrease its frame transmission rate over-aggressively, and then operate with a lower transmission rate than the actual

achievable higher rate. Second, since each contending station conducts its rate adaptation independently, they may end up with transmitting data at different rates. Such transmission rate diversity causes the performance anomaly that was first discovered experimentally in [15]. Since the 802.11 DCF is designed to offer equal transmission opportunities (or long-term equal medium access probabilities) to all contending stations, the throughput of a high-rate station is always bounded below the lowest transmission rate in the network.

6.3 Fully Connected Topologies with Varying Number of Contending Stations and Time Varying Channel

Fig. 5 shows the performance with fading effect. The simulation parameters are same as that of Fig. 4. Ricean fading with line-of-sight factor $K = 0$ is considered. The saturated throughput is obviously lower than the earlier case because of fading effect. So the offered load in this case is 2 Mbps. It is to be noted that the coherence time assumed is sufficiently large for at least two packets transmission at the same data rate. The system estimates the channel for every two consecutive successful transmissions (m_{thr}).

It is observed that the performance of ARF is very poor even for a small network size of 4 nodes. The aggregate throughput is degraded up to 0.5 Mbps. Even the aggregate throughput of CARA is going below 1 Mbps. This may be because of the time varying channel. The time varying nature of the channel is well understood by the SNR values of the received signal and accordingly the data rate is selected. Effectively TCECLD performs better and achieves significantly higher, aggregated system throughput than CARA and ARF in this simulation setup.

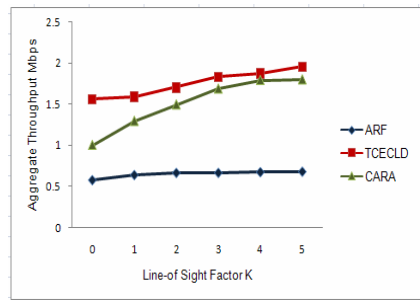
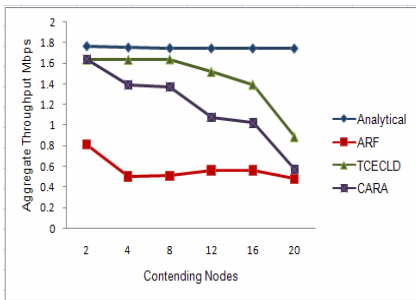


Fig. 5. Aggregate Throughput as a function of Contending Nodes in Time Varying Channel Conditions

Fig. 6. Aggregate Throughput as a function Line-of-Sight Factor K

6.4 Effect of Line of Sight Component K

Here, the effect of the Ricean parameter K is explored on the performance of ARF, CARA and TCECLD. For $K = 0$, the channel has no line-of-sight component such that only reflected signals are received and hence, overall channel quality is poor. With increasing K , the line-of-sight component is stronger such that the overall

channel SNR increases as described by Equation (3), and a higher transmission rate is feasible more often.

Fig. 6 depicts the aggregate throughput for ARF, CARA and TCECLD as a function of the Ricean parameter K for a fully connected 10 node random topology. Observe that both CARA and TCECLD exploit the improved channel conditions represented with increasing K and obtain correspondingly greater system-wide throughputs. Moreover, note that TCECLD achieves a higher aggregate throughput compared to ARF and CARA over the simulated range of K due to its enhanced exploitation of high-quality channel conditions when they occur.

7 Conclusion

In this paper, a novel Transmitter based Capacity Enhancement algorithm with Cross-Layer Design approach in wireless ad-hoc network is proposed. The key idea is that the transmitter station estimates the channel condition and accordingly selects the data rate for transmission. The parameter transmission data rate is tuned to exploit the channel conditions. Therefore, compared with ARF, the most well-known and widely-deployed rate adaptation scheme in the commercial 802.11 WLAN devices, it is more likely to make the correct rate adaptation decisions. Moreover, no change is required in the current IEEE 802.11 standards, thus facilitating its deployment with existing 802.11 devices.

The performance is evaluated via in depth simulations over various scenarios in terms of network topology, offered load and time varying wireless channel. It is demonstrated that the proposal significantly outperforms ARF in all the simulated multiple contending station environments, whereas the performance enhancement becomes more and more evident as the number of contending stations increases.

References

1. Kamerman, A., Monteban, L.: WaveLAN II: A High Performance Wireless LAN for the Unlicensed Band. Bell Labs Technical Journal, 118–133 (Summer 1997)
2. Holland, G., Vaidya, N., Bahl, P.: A Rate Adaptive MAC Protocol for Multi-hop Wireless Networks. In: Proc. ACM MOBICOM 2001, Rome, Italy, pp. 236–251 (2001)
3. IEEE 802.11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications (June 1997)
4. Rappaport, T.S.: Wireless Communications: Principles and Practice. Prentice Hall, Englewood Cliffs (1999)
5. Chevillat, P., Jelitto, J., Noll Barreto, A., Truong, H.L.: A Dynamic Link Adaptation Algorithm for IEEE 802.11a Wireless LANs. In: Proc. IEEE ICC 2003, Anchorage, AK, pp. 1141–1145 (May 2003)
6. del Prado Pavon, J., Choi, S.: Link Adaptation Strategy for IEEE 802.11 WLAN via Received Signal Strength Measurement. In: Proc. IEEE ICC 2003, Anchorage, AK, vol. 2, pp. 1108–1113 (May 2003)
7. Sadeghi, B., Kanodia, V., Sabharwal, A., Knightly, E.: OAR: Opportunistic Auto Rate Media Access Protocol for Ad Hoc Networks. IEEE/ACM Trans. on Wireless Networks 11(1-2), 39–53 (2005)

8. Qiao, D., Choi, S., Shin, K.G.: Goodput Analysis and Link Adaptation for IEEE 802.11a Wireless LANs. *IEEE Trans. on Mobile Computing (TMC)* 1(4), 278–292 (2002)
9. Qiao, D., Choi, S.: Fast-Responsive Link Adaptation for IEEE 802.11 WLANs. In: *Proc. IEEE ICC 2005*, Seoul, Korea, vol. 5, pp. 3583–3588 (May 2005)
10. Kim, J., Kim, S., Choi, S., Qiao, D.: CARA: Collision-Aware Rate Adaptation for IEEE 802.11 WLANs. In: *Proc. IEEE INFOCOM 2006*, Barcelona, Spain, pp. 1–11 (April 2006)
11. Shakkottai, S., Rappaport, T.S., Karlsson, P.C.: Cross-Layer Design for Wireless Networks. *IEEE Comm. Magazine* 41(10), 74–80 (2003)
12. Conti, M., Maselli, G., Turi, G.: Cross-Layering in a Mobile Ad-hoc Network Design. *IEEE Comp. Soc.* 37(2), 48–51 (2004)
13. Bianchi, G.: Performance Analysis of the IEEE 802.11 Distributed Coordination Function. *IEEE Journal on Selected Areas in Communications* 18(3), 535–547 (2000)
14. Punnoose, R.J., Nikitin, P.V., Stancil, D.D.: Efficient Simulation of Ricean Fading within a Packet Simulator. In: *Proc. IEEE VTS-Fall VTC 2000*, vol. 2, pp. 764–767 (September 2000)
15. Heusse, M., Rousseu, F., Berger Sabbatel, G., Duda, A.: Performance Anomaly of 802.11b. In: *Proc. IEEE INFOCOM 2003*, vol. 2, pp. 836–843 (March 2003)

Wireless Sensor Network Using Bluetooth

Omkar Javeri and Amutha Jeyakumar

Department of Electronics Engineering, Veermata Jijabai Technological Institute,
Mumbai, India

omkar_javeri@yahoo.com, amuthajaykumar@vjti.org.in

Abstract. A Wireless Sensor Network (WSN) is an ad-hoc wireless network formed of spatially distributed autonomous sensor nodes which possess the ability of performing computations, communication & executing different sensing tasks. The design & development of wireless sensor nodes is carried out using Bluetooth as wireless networking technology. The use of Bluetooth helps to utilise many features of Bluetooth enabled devices, which could be used as an infrastructure gateways, actuators or user interfaces. A modular design approach is adapted for design & development of wireless sensor node thus making it easy to add, remove or modifying the sensor with much shorter development cycles while also supporting WSN-behavior in terms of power consumption & node operational lifetime. WSN architecture supports different network topologies like Star, mesh & hybrid Star-mesh network. An embedded wireless sensor node is designed & successfully tested for two different Star network formations.

Keywords: Bluetooth Technology, Wireless Sensor Node & Network Architecture, Wireless Sensor Node design, WSN formation.

1 Introduction

A Wireless Sensor Network (WSN) is an ad-hoc network formed by spatially distribution of wireless sensor nodes to form a cooperative network. These nodes possess the ability of performing computations; communication & executing different sensing tasks by cooperatively monitor different physical or environmental conditions, such as temperature, sound, vibration, pressure, motion or pollutants, at different locations. Wireless sensor networking makes use of advance wireless integrated circuits to couple wireless subsystems to sophisticated sensors which consist of one or more micro-controllers, a RF transceiver, memory, a power source & contain various types of sensors & actuators. The nodes are autonomous & communicate wirelessly after being deployed in an ad-hoc fashion. Systems of 100s of nodes are anticipated. Such systems will transform the way we live & work [1].

The current generation of sensor nodes relies on commodity components. The choice of the radio is particularly important as it impacts not only energy consumption but also software design (e.g., network self-assembly, multi-hop routing & in-network processing). The Bluetooth wireless radio which works in 2.4 GHz ISM band is a good choice for implementation of wireless sensor network. The Wireless Sensor

Networks architecture supports different network topologies like Star, mesh & hybrid Star-mesh network.

The embedded wireless sensor node is designed which consist of AT mega 128 microcontroller, Parani-ESD 100 Bluetooth module, two different signal sensing & conditioning circuits for analog & digital signals obtained from different sensors, LCD & regulated power supply. Different sensor circuits are used to sense motion, temperature, vibration, acoustic sound & light intensity & is used to monitor environmental conditions, N number of such nodes are developed along with different daughter boards for analog & digital sensor separately. The embedded node is further successfully tested for two different Star network formations.

2 The Bluetooth Technology

Bluetooth™ is the codename for a technology specification for low cost, short range radio links between personal computer, mobile phones & other portable devices. Unlike many other wireless standards, the Bluetooth wireless specification includes both link layer & application layer. Radios that comply with the Bluetooth wireless specification operate in the unlicensed, 2.4 GHz radio spectrum ensuring communication compatibility worldwide. Bluetooth radios use a frequency hopping spread spectrum, full duplex signal at up to 1600 hops/sec. The signal hops among 79 frequencies at intervals of 1 MHz to give a high degree of interference immunity. The Bluetooth specification contains the information necessary to ensure that diverse devices supporting the Bluetooth wireless technology can communicate with each other worldwide [2].

2.1 The Bluetooth Specification and Power Classes

Bluetooth Specification version 1.1 consists of Core Specification (Volume I) which defines radio & hardware oriented software protocols to ensure interoperability between devices from different manufacturers. The Profile Definitions (Volume II) describe standardized software applications defining messages & procedures. There are 13 approved Bluetooth profiles illustrating a variety of Bluetooth applications [3]. Each device is classified into 3 power classes, Power Class 1, 2 & 3 designed for long range (~100m), ordinary range (~10m) & short range devices (~1m) devices, with a max output power of 20dBm, 4dBm & 0dBm respectively.

2.2 Piconet

Piconet consists of a collection of Bluetooth enable devices connected in an ad-hoc fashion. A piconet consists of minimum two connected devices, such as a portable PC & cellular phone, & may grow up to eight connected devices. All Bluetooth devices are connected as peer to peer units & have identical implementations. In order to establish a piconet, one device will act as a master & the other devices as slave for a particular piconet formation & specific duration of time.

3 Wireless Sensor Node and Network Architecture

3.1 Wireless Sensor Node Architecture

The Wireless Sensor Node consists of a microprocessor module, a radio module like wireless Bluetooth module, signal sensing circuit, display & regulated power supply. The microprocessor performs the majority of functions like managing data collection from the sensors, performing power management functions, interfacing the sensor data to the physical radio layer & managing the radio network protocol [4, 5].

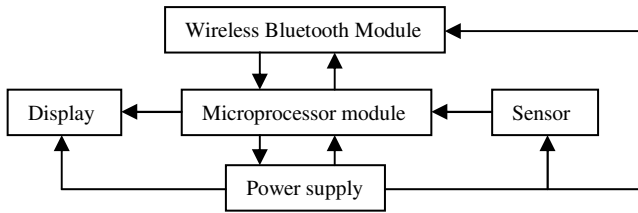


Fig. 1. Wireless sensor node architecture

A modular design approach provides a flexible & versatile platform to address the needs of a wide variety of applications [5]. For example, depending on the sensors to be deployed, the signal conditioning block can be re-programmed or replaced. This allows for a wide variety of different sensors to be used with the wireless sensing node. Similarly, the radio link may be swapped out as required for a given applications, wireless range requirement or the need for bidirectional communication. The use of flash memory allows the remote nodes to acquire data on command from a base station, or by an event sensed by one or more inputs to the node.

One of the most important features of any wireless sensing node is to minimize the power consumed by the system. Generally, the radio subsystem requires the largest amount of power. Therefore, it is advantageous to send data over the radio network only when required using a suitable algorithm. Additionally, it is important to minimize the power consumed by the sensor itself [6]. Therefore, the hardware should be designed to allow the microprocessor to judiciously control power to the radio, sensor, & sensor signal conditioner.

3.2 Wireless Sensor Networks Architecture

There are a number of different topologies supported for wireless network formation such as Star, Mesh & Hybrid Star-Mesh Network. In Star Network single node acts as the network coordinator & is responsible for network formation & control. In Mesh network every node is connected to other to form a mesh which allows any node in the network to transmit to any other node thus allowing multi-hop communications. It has an advantage of redundancy & scalability but consumes more power as compare to Star. A Hybrid network formed out of Star & Mesh network provides for a robust & versatile communications network, while maintaining the ability to keep the wireless sensor nodes power consumption to a minimum [7].

4 Wireless Sensor Node Design

The wireless sensor node is built around an Atmel ATmega128L microcontroller with on-chip memory & peripherals. The microcontroller features 8-bit RISC cores with capability of performing up to 8 MIPS at a maximum of 8 MHz. The on-chip memory consists of 128 Kbytes of in-system programmable Flash memory, 4 Kbytes of SRAM, & 4 Kbytes of EEPROM. There are several integrated peripherals: JTAG for debugging, timers, counters, pulse-width modulation, 10-bit analog-digital converter, I2C bus, & two hardware USARTs. The operating voltage ranging from 2.7 – 5.5V thus giving a wide range of flexibility & is suitable fulfils with the requirements of other low voltage devices working at 3.3 V.

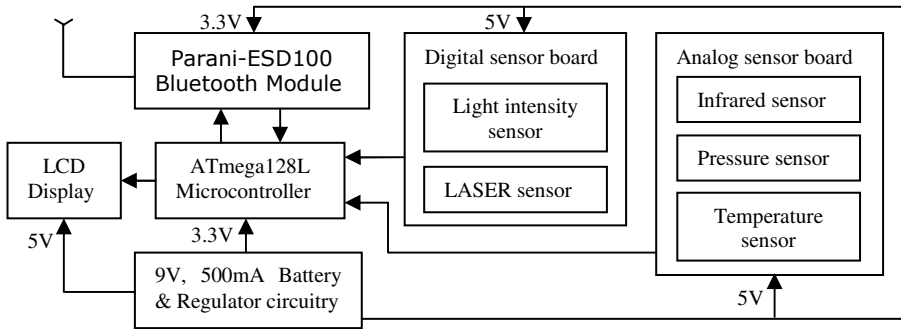


Fig. 2. Wireless sensor node design

A Parani-ESD 100 Bluetooth module is used to setup the wireless sensor link. Parani-ESD100 Bluetooth module is connected to one of the serial ports of the microcontroller using a detachable module 12 pin connector. Six LEDs are integrated, mostly for the convenience of debugging & monitoring. The entire board is powered using a 9V battery after the voltage being regulated to 5V & 3.3V as per the requirement. Connectors that carry both power & signal lines are provided & can be used to add external peripherals, such as sensors & actuators.

4.1 Parani-ESD

Parani-ESD is a module device for wireless serial communication using Bluetooth technology. It can communicate with other Bluetooth devices that support the Serial Port Profile. Parani-ESD100/110 module has a communication range of 100/100-1000 meter in open space. The default range is 100m using default antenna & the range can be extended to 1000m using patch-patch antenna. It can be configured & controlled by typical AT commands by using a terminal program such as HyperTerminal & can use Bluetooth wireless communication without modifying user's existing serial communication program. In addition to the basic AT commands, some expanded AT commands are used for various functions [8, 9].

4.2 ATmega128L

The ATmega128L is a low-power CMOS 8-bit microcontroller based on the AVR enhanced RISC architecture. By executing powerful instructions in a single clock cycle, the ATmega128L achieves throughputs approaching 1 MIPS per MHz allowing the system designed to optimize power consumption versus processing speed [10].

4.3 AT Command

AT command set is a standard language for controlling modems. AT is the abbreviation of ATtention. Every command line starts with prefix "AT". The AT command set was developed by Hayes & is recognized by virtually all personal computer modems [11]. Parani-ESD provides the extended AT command set to control & configure the serial parameters & Bluetooth connection. It replies to AT commands with 4 message, 'OK', 'ERROR', 'CONNECT' & 'DISCONNECT' [9].

4.4 Configuration and Operation Modes of Parani-ESD

In addition to the serial port configurations, the Parani-ESD also requires some settings for Bluetooth. A Bluetooth connection is always made by a pair of master & slave devices. Master tries to connect itself to other Bluetooth devices, & slave is waiting to be connected from other Bluetooth devices. A slave can be in two modes, Inquiry Scan or Page Scan mode. Inquiry Scan mode is waiting for a packet of inquiry from other Bluetooth device & Page Scan mode is waiting for a packet of connection from other Bluetooth device. Every Bluetooth device has its unique address, called Bluetooth Device address, which is composed of 12 hexadecimal numbers [8, 9].

4.5 Hardware Flow Control

The Bluetooth module temporarily stores the data into its internal buffer & sent repeatedly until the transmission is completed packet by packet. A suitable Hardware Flow Control (HFC) is used by the Bluetooth module. When the radio transmission condition is not good enough to send data promptly, it can cause a transmission delay. If host sends more data when the buffer is full, buffer overflow will make Bluetooth module malfunction consequently. To prevent this buffer overflow when using HFC, Bluetooth module disables RTS so that it stops receiving any further data from the host when the buffer becomes full. RTS will be re-enabled again to begin receiving data from the host when the buffer has created more room for more data [8], [9].

4.6 Analog and Digital Signal Sensing Circuits

Analog signal sensing circuit is designed & printed on a different PCB it consist of four different analog sensors circuits for detecting motion, pressure sensed, sudden rise in temperature & to detect acoustic sound. Similarly Digital signal sensing circuit is designed & printed on a different PCB it consist of two different digital sensors circuits for detecting light intensity & LASER light.

5 Wireless Sensor Network Setup

Wireless Sensor Network setup consists of node development phase where sensor motes are placed at different location in the surrounding where the entity of interest is to be observed. It consists of ATmega128L microcontroller at the heart of the sensor mote which manages data collection from different sensors. These sensors are embedded in another PCB & are connected to the sensor mote using detachable cables. These nodes are future configured to form a Star network and enter in sleep mode when ideal for long time to conserve power.

5.1 Star Network Formation with Computer Acting as the Host Controller and Sensor Motes as End Point Devices

The Star network formations consist of a personal computer acting as the host controller & several sensor motes connected to the host controller as end point devices. The configuration for initialization of Bluetooth module is written in the configuration program placed in the microcontroller where the baud rate, parity, stop bit are set also the mode of operation & hardware flow control setting are set. The host controller is responsible for establishing the Bluetooth connection with the sensor motes. The personal computer is equipped with a Bluetooth dongle as the Bluetooth radio hardware. The personal computer uses Blue-Soleil software to search for Bluetooth signal from other Bluetooth devices in its vicinity [12]. The sensor mote Bluetooth modules are set in mode 3 so as Parani-ESD is discoverable & is waiting for the connection from any other Bluetooth devices. After Blue-Soleil completes its search for Bluetooth signals list of Bluetooth devices in the vicinity of 1000 meter are displayed. The Parani-ESD motes displayed is paired one by one & a suitable serial communication link is established between the personal computer & the sensor motes to form a Star network. At a time a maximum of seven motes be simultaneously connected to the host controller to form a Star network. In order to connect more than seven motes one of the currently connected motes is disconnected & then the connection with other mote is established. For more secure Bluetooth connection the Authentication & data Encryption option can be used where a secret pass key is required before establishing a connection between the host controller & the sensor mote. The data transmitted over the Bluetooth channel is encrypted before transmission & is received as noise by others who try to hack the signal.

After the Star Bluetooth link is successfully established. The microcontroller processes the data obtained from its different sensors & send it to the Bluetooth module for transmission. The Bluetooth module temporarily stores the data into its internal buffer & sent repeatedly until the transmission is completed packet by packet. A Hyper terminal program is used to acquire the serial data transmitted by the Bluetooth module. A terminal window is opened for every node connected in Star network where particular serial port is selected & setting of baud rate, parity, stop bit, & hardware flow control is entered this is in accordance with the particular node selected & the setting for that node which was placed in the configuration program for node in the microcontroller. The respective data send by the sensor motes is displayed on the each terminal window & can be continuously monitored. The motes connected now can accept AT commands when written in the terminal window. A suitable data log

file is created for the entire transmitted & received data & the AT commands typed. The data being received can be used to plot a graphical view of the output characteristics of the different sensors being used with the help of software such as Visual Basic & MATLAB.

5.2 Star Network Formation with One Mote Acting as Host Controller and Other Motes Acting as Endpoint Devices

The Star network formations consist of a sensor mote acting as the host controller & several other sensor motes connected to the host controller as end point devices. The configuration for initialization of Bluetooth module is written in the configuration program placed in the microcontroller where the baud rate, parity, stop bit are set also the mode of operation & hardware flow control setting are set. The host controller mote is responsible for establishing the Bluetooth connection with other sensor motes. The sensor mote Bluetooth modules are set in mode 3 so as Parani-ESD is discoverable & is waiting for the connection from any other Bluetooth devices. When the host mote is ready to transmit the processed data obtained from its different sensors then it scans for the other Bluetooth enable nodes which are in a radius of 100 meters from it. The host node gets a list of Bluetooth address of the other nodes now it makes a connection with a node to which it wants to transmit the data. Once the connection has been established then the data is send to the Bluetooth module for transmission. The Bluetooth module temporarily stores the data into its internal buffer & sent repeatedly until the transmission is completed packet by packet. Once the transmission is completed the host disconnects the currently connected device to conserve power as maximum power is used during data transmission. In order to send the data to another node it establishes as connection with that node & then transmits the data & disconnects the current connection. It takes 5 seconds to establish a new connection the connection time may increase depending on the number of devices in its range, interference by other devices & the surrounding environmental conditions. It is possible for host node to form a Star network with N number of nodes & not restricted to only seven nodes as in previous scenario thus making it possible to place as many nodes in Star fashion as required by the user but the load on the host node increases as the number of node connected in Star fashion increases.

5.3 Application in Forest Monitoring

The Wireless sensor nodes are configured to form a wireless network where each node is connected to network coordinator in a Star fashion which act as a base station where the entire data is collected & monitored form. Wireless Sensor Network using these sensor nodes was used to monitor the presence of animals & monitor different parameters such as acoustic sounds, temperature & pressure without affecting the natural behaviour of animals. The issue of forest fire is common in many countries. The sensor is also configured to detect the presence of fire which after detection sends the alarm signal to the base station. This is very useful to detect forest fire where remote monitoring can help in early detection of fire & necessary action be taken in

an early stage. The Wireless Sensor Network is also useful to monitor the presence of intruder in the forest. This would help in to take care of unlawful actions like hunting of animals & cutting of trees in the forest.

6 Conclusion

The embedded wireless sensor mote developed using a Bluetooth as wireless networking technology help to incorporate different features of Bluetooth. Bluetooth modules built fully to specification are well suited for sensor networks. With data rates up to 1 Mbps, Bluetooth also offers more than enough bandwidth for simple sensor networks.

The embedded node developed is very compact & possesses the ability of performing computations & communication, executing different sensing tasks. The modular design approach has resulted into flexible & versatile platform to address the needs of a wide variety of applications. The design also provides a feature to easily add, modify or replace any sensors as per the need of the application & the user. Prototype built has been successfully tested for different network formations with a personal computer acting as a network coordinator & further the sensor motes were successfully programmed to act as the network coordinator.

A successful Star network is established with personal computer acting as the network coordinator connecting wireless sensor nodes. A successful transmission of different set of data obtained from different sensors after signal processing & conditioning by the sensor nodes is transmitted from the sensor node to the host controller. The data reception is continuous & is obtained in real time. The transmission is successfully recorded for long hours without any link getting broken. The dynamic changes in the position of node in the radius of 100m from the host controller in the middle of transmission did-not affect the link quality & the transmission was continuous. The algorithm developed to use the hardware flow control feature of the Bluetooth module & to save power by placing a particular sensor node in sleep mode when not in use help to considerably enhance the life of battery operated nodes.

The Star network formed with one sensor mote acting as the network coordinator is able to perform successful data transmission of data with other sensor nodes. The host node is program to establish a connection only when it require to transmit the data so as to conserve maximum power since maximum power is consumed during data transmission. It was also observed that as the number of devices connected to the host controller node increases the load on the host controller increases & thus degrading the performance of the network when many nodes want to transmit data simultaneously.

Wireless sensor networks are thus enabling applications that previously were not practical. With new standards based networks being released & low power systems are continually developed, we will start to see the widespread deployment of wireless sensor networks. In future we will see rise in use of Wireless sensor networks in a wide spread area of different applications.

7 Future Scope

One issue that needs further investigation is the scalability of Bluetooth sensor networks. The scatter-net support in the Bluetooth specification is not suitable for multi-hop networks today. One approach that could enable this consists of adding support for a broadcast channel, which could be used as a wake-up radio & also enable formation of multi-hop networks. The power consumption issue of the Bluetooth radio module also needs to be addressed. The development of the next generation of Bluetooth modules with an additional ultra low-power radio & the use of dual powered sensor nodes with power from combination of rechargeable batteries, super capacitors & solar panels which charge the batteries during the day & can be used throughout the day will allow to build more energy efficient sensor networks based on Bluetooth thus combining the energy efficiency of ultra low power radios with the interoperability of Bluetooth with long operational life time.

References

1. Stankovic, J.A.: Wireless Sensor Networks, Department of Computer Science, University of Virginia Charlottesville, Virginia (June 2006)
2. The Official Bluetooth Technology Info Site (August 2010), <http://www.bluetooth.com/>
3. Rodzevski, A., Forsberg, J., Kruzela, I.: Wireless Sensor Network with Bluetooth, Intelligent Systems Group School of Technology & Society, University of Malmö, Sweden
4. Beutel, J., Kasten, O., Mattern, F., Romer, K., Siegemund, F., Thiele, L.: Prototyping Wireless Sensor Network Applications with BTnodes, Institute for Pervasive Computing Department of Computer Science ETH Zurich, Switzerland
5. Leopold, M., Dydensborg, M.B., Bonnet, P.: Bluetooth & Sensor Networks: A Reality Check, Department of Computer Science. University of Copenhagen,
6. Luleå, J.E.: Low-Power Design Methodologies for Embedded Internet Systems, University of Technology Department of Computer Science & Electrical Engineering, EISLAB, 2008, pp. 3–16, 23–31 (2008)
7. Wilson, J.: Sensor Technology Handbook Book, pp. 439–448. Elsevier Science & Technology Books (December 2008)
8. Sena Technologies, Inc., Bluetooth Serial Module, http://www.sena.com/products/industrial_bluetooth/esd100.php
9. Sena Technologies, Inc., Parani-ESD100/110 User Manual, http://www.sena.com/download/manual/manual_parani_esd-v1.1.6.pdf
10. Atmel Corporation, Atmel AVR 8-Bit RISC (2010), <http://atmel.com/products/avr/default.asp>
11. Wikipedia, Hayes command set, http://en.wikipedia.org/wiki/Hayes_command_set
12. IVT Corporation, Bluesoleil, <http://www.bluesoleil.com/Default.aspx>

A Performance of Security Aspect in WiMAX Physical Layer with Different Modulation Schemes

Rakesh Kumar Jha¹, Suresh Limkar², and Upena D. Dalal¹

¹Department of Electronics and Communication Engineering, SVNIT Surat, India

²Department of Computer Engineering, SVNIT Surat, India

Jharakesh.45@gmail.com, sureshlimkar@gmail.com,

upena_dalal@yahoo.com

www.jharakeshnetworks.com

Abstract. This paper presents WiMAX physical layer threats jamming and scrambling. The performance of the system was found out to greatly differ with the use of different jamming signals, allowing central areas to be identified, where system development should be focused on. In addition, from the basic theory point of view, rather surprising results were also found. The work should give a clear picture of how the studied WiMAX system performs under jamming as well as without the presence of jamming. The results show that some forms of interference degrade the performance of the system rapidly, thus the form of incoming jamming should be known and considered before deploying the system. Noise jamming, multi-carrier jamming and scrambling are discussed here. The issues related to jamming and jamming reduction techniques are also covered. Jamming and scrambling can destroy communication in the targeted area. Multi-carrier jamming is challenge in WiMAX because WiMAX is having OFDM based physical layer. Simulation approach is main concern here. OPNET MODELER is the software used for the simulation purpose.

Keywords: WiMAX, Jamming, Scrambling.

1 Introduction

IEEE 802.16 is the standard for WiMAX. WiMAX is also known as wireless broadband. IEEE 802.16-2004 is known as fixed WiMAX and IEEE 802.16-2005 is known as mobile WiMAX [1]. In wired networks physical layer threats are not important but in wireless air is used as medium so physical layer threats comes into picture. In wireless jamming and scrambling are considered as physical layer threats. Here simulation approach is used to see the performance of the IEEE 802.16-2004 system in jamming and scrambling environment [8], [10]. Jamming is achieved by introducing a source of noise strong enough to significantly reduce the capacity of the WiMAX channel. The information and equipment required to perform jamming are not difficult to acquire. Resilience to jamming can be augmented by increasing the power of signals or increasing the bandwidth of signals via spreading techniques such as frequency hopping or direct sequence spread spectrum. The practical options

include a more powerful WiMAX transmitter, a high gain WiMAX transmission antenna, or a high gain WiMAX receiving antenna. It is easy to detect jamming in WiMAX Communications as it can be heard by the receiving equipment. Law enforcement can also be involved to stop jammers. Since jamming is fairly easy to detect and address, so it does not pose a significant impact on both the WiMAX users and systems. Scrambling is usually instigated for short intervals of time and is targeted to specific WiMAX frames or parts of frames [3] [9]. WiMAX scramblers can selectively scramble control or management messages with the aim of affecting the normal operation of the network. Slots of data traffic belonging to the targeted SSs can be scrambled selectively, forcing them to retransmit. Noise jamming and multi-carrier jamming are considered here for simulation approach. Noise jamming is used to jam the particular band of frequencies. In noise jamming carrier frequency and bandwidth of the targeted system should be known. In multi-carrier jamming the frequencies of carriers of targeted system should be known. Simulation approach is easy compare to practical approach. The issues related to practical approach will be described in the later part.

2 The Investigated Physical Layer

The primary operation bands of IEEE 802.16-2004 include 10-66 GHz licensed bands, frequencies below 11GHz and license-exempt frequencies below 11GHz (primarily 5-6 GHz) [1]. According to these operation bands, IEEE 802.16-2004 PHY defines five specifications for different operation scenarios. Among them, Wireless MAN-OFDM PHY is based on orthogonal frequency-division multiplexing (OFDM) technology and designed for NLOS operation in the frequency bands below 1GHz. It is selected to be the air interface of the system under investigation in this paper.

At the transmitter side, the information data first undergoes channel coding composed of randomization, forward error correction (FEC), and interleaving. Randomizer uses a Linear Feedback Shift Register (LFSR) to scatter long data strings of zeros or ones. Forward error correction concatenates an outer Reed-Solomon encoder with an inner rate compatible convolutional encoder. FEC helps to correct the errors in subcarriers to a certain limit. The interleaver takes two permutations to rearrange the subcarriers so that the burst errors are distributed more uniformly at the demodulation input [2]. After channel coding, data bits are mapped and modulated onto the allocated subcarriers by BPSK, 16-QAM and 64-QAM modulation. Subsequently, data are transmitted by OFDM method. In the receiver side, all the procedures carried out in the transmitter side are implemented again but in a reverse direction. One OFDM symbol can be divided into two parts in time domain: the cyclic prefix (CP) time and the useful symbol time. The cyclic prefix locates in the beginning of the symbol and is a duplication of the tail of the useful symbol, which is introduced to mitigate the effect of multipath. In frequency domain, an OFDM symbol is composed of a series of subcarriers. In Wireless MAN-OFDM PHY, the number of subcarriers is 256. As shown in Fig. 1, three types of subcarriers can be categorized: 192 data subcarriers carrying payload, 8 pilot subcarriers mainly for channel estimation, and 56 null subcarriers for guarding purpose. The pilot subcarriers distribute evenly among the data subcarriers. This is standard symbol in frequency domain.

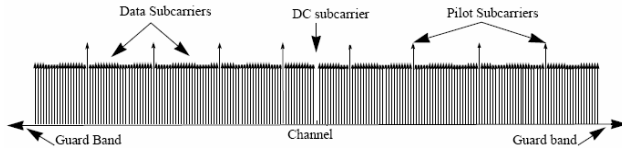


Fig. 1. OFDM symbol Frequency domain description

Channel estimation is mandatory for the OFDM systems employing coherent detection. Comb type pilot channel estimation is capable of collecting instant information of the channel and therefore used in this research. The channel estimation for the payload subcarriers is achieved by interpolation, using the channel information obtained at the 8 pilot subcarriers.

3 Simulation

Simulation can be done in any software but the procedure remains same. For simulation some parameters are taken from the standard and some parameters are varied to get the appropriate results. Modulation parameters used in the simulation are listed in the Table.1

Table 1. Parameters Used in Simulation Algorithm [7]

Modulation	Code rate for convolution coding
BPSK	1/2
QPSK	1/2
QPSK	3/4
16-QAM	1/2
16-QAM	3/4
64-QAM	2/3
64-QAM	3/4

The computer simulation in this paper is generated using OPNET MODELER. Besides jamming, the system in the simulation is subjected to multipath fading and additive white Gaussian noise. The multipath channel is simulated as a frequency selective, slow fading channel by snapshot method. The parameters adopted in the simulation are listed in Table.2.

The values of parameters are taken from standards decided by IEEE 802.16. All this values are considered for the simulation and remains same for number of scenarios. For comparison of scenarios to be simulated the selection of values is important. Noise jamming and multi-carrier jamming can be simulated if system bandwidth and carrier bandwidth is known. But in practical there are so many issues related to antennas which will be discussed later.

Table 2. Physical Layer Parameters and Their Values for Simulation [7]

Parameters	Standard value for simulation
Channel bandwidth	20 MHz
Number of carriers	200
CP ratio	1/4
OFDM symbol duration	102 μ s
Number of FFT points	256
Sampling factor n	28/25
SNR	20 db

4 Mathematics Involved

There are several QOS (quality of service) parameters to evaluate the system performance under jamming. The formula given below is useful to define jamming type.

$$\frac{B_j}{B_{vs}} = \frac{\text{Jammerbandwidth}}{\text{Victimsystembandwidth}} \quad (1)$$

If the ratio B_j/B_{vs} is less than 0.2 jamming is considered to be spot (narrow band) jamming and if greater than 1, barrage (wide band) jamming [7]. Packet error rate and signal to jamming ratio is used to evaluate the system performance. There are several other QOS parameters like throughput, delay and traffic related parameters that can be used to evaluate the system performance. The packet error rate can be calculated by the formula given below.

$$PER = \frac{\text{erronouspacket}}{\text{packetsent}} \quad (2)$$

The measurement was conducted by transmitting constant length (8 kb) UDP (User Datagram Protocol) packets over the connection, with a constant transmission rate of 95 % of the measured maximum throughput allowed by the selected modulation/coding combination. The transmission rate was selected 5 % lower than the maximum to make sure that no errors occur because of the small fluctuations in the system capacity caused by the software, computers, network adapters' etc. signal to noise ratio is given by the following equation. Here noise power is jamming signal power.

$$SNR = \frac{\text{Signalpower}}{\text{Noisepower}} \quad (3)$$

5 Design of Jammer in OPNET Modular

The transmitter node model of jammer is given in the figure.

The tx_gen is processor module which calculates the information that the antenna needs to point at a target: latitude, longitude, and altitude coordinates. The pointing

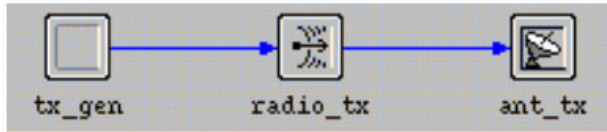


Fig. 2. Transmitter node model of jammer

processor makes this calculation by using a Kernel Procedure that converts a node’s position in a subnet (described by the x position and y position attributes) into the global coordinates that the antenna requires.

The radio_tx is radio transmitter module which transmits packets to the antenna at 1024 bits/second, using 100 percent of its channel bandwidth. For each arriving candidate packet, the radio receiver module consults several properties to determine if the packet’s average bit error rate (BER) is less than a specified threshold. If the BER is low enough, the packet is sent to the sink and destroyed.

The ant_tx is antenna module which models the directional gain of a physical antenna by referencing its pattern attribute. The antenna uses two different patterns: the isotropic pattern (which has uniform gain in all directions) and a directional pattern. Antenna pattern editor is also provided in the software. Antenna pattern is important in the design of jammer. By using this software antenna pattern can be designed to utilize power effectively.

Introduction to design of jammer is given in this paper. It is not possible to include step by step procedure. The idea is given here how to design jammer using opnet modeler.

6 Results

Results are generated by the software simulation. Performance of the IEEE 802.16 system is measured in terms of the QOS parameters. With out jamming throughput is

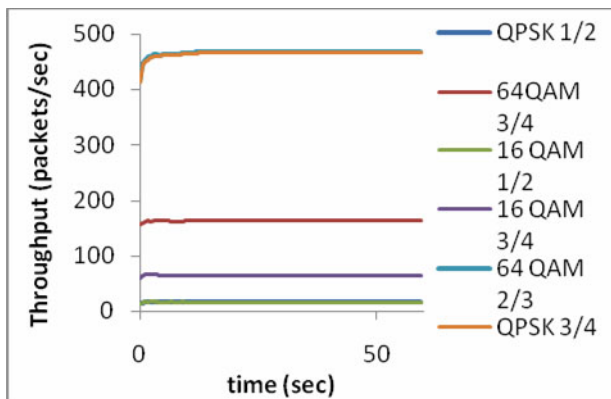


Fig. 2. Result of video conferencing application using WiMAX for period of one minute without jamming

maximum (97%) and packet error rate is minimum. The result of video conferencing is shown below. Video conferencing application is simulated using WiMAX for one minute with parameters defined in simulation section.

With jamming the throughput of the system decreases and packet error rate increases. Jamming power required for certain values of PER and specific modulation schemes are given below. Modulation and simulation parameters are described in simulation section.

Table 3. Jamming power (dB) required for specific modulation scheme and PER in downlink

PER	QPS K 1/2	QP SK 3/4	16 QAM 3/4	64 QAM 2/3	64 QAM 3/4
5%	-35.8	- 40.1	-43.8	-42	-43.3
30%	-34.8	-39	-42.7	- 40.6	-41.9
60%	-33.8	- 38.5	-41.9	- 39.8	-41.2

The result shows that jamming power increases for higher values of PER for particular modulation scheme. As jamming power is increased PER also increases so throughput decreases. Signal to noise ratio decreases as the jamming power increases. Simulated results give idea about the practical scenario. Practical results may be different from simulated results but concept remains same.

7 Issues and Overview of Jamming Reduction Techniques

There are number of issues related to practical approach. Simulation approach is easy but not reliable because the results depend on the software and its limitation. In simulation some results may not be generated but it is cost effective solution. In practical there are so many issues related to antennas used to produce jamming signal. The placement of jamming antenna and its polarization also affects in practical case. Scrambling is very difficult to simulate because it targets the part of packet [3]. Practical approach is very useful to perform scrambling. Smart antennas are not included in the software but it can be used practically. Polarization of the victim system antenna and jamming antenna is not matched then it results in the loss of power. Some practical parameters cannot be included in software. There are several techniques to reduce the jamming.

7.1 Beamforming

It is the technique to improve the directionality of the antenna. This technique is useful in the base station which uses sector pattern instead of the Omni directional antenna. Let's consider three sector of base station antenna. If jamming antenna is in sector one then the other two sectors are less affected by the jammer. This technique

is useful in reducing jamming effect but it cannot remove it. This technique is very difficult to implement because reduction of the main lobe of antenna is complicated.

7.2 Jamming Resistant Architecture

This jamming resistant architecture is also used to reduce the jamming effects [6]. In this architecture more than one base station is used in one cell. All subscriber stations are connected in mesh topology including base stations which is serving in particular cell. Let's consider two base stations in one cell. If jamming is introduced and one base station fails then the all subscriber stations served by that base station are automatically connected to other base station. There is one major problem in this architecture related to scheduling. Scheduling is the mechanism which is used by the base station to assign the resources to the subscriber stations. There are more than one base stations in one cell so distributed scheduling must be used. Distributed scheduling is very complicated so it is very difficult to design it. Scheduling is the mechanism which is used by the base station to assign the resources to the subscriber stations.

7.3 OFDM-CDMA

OFDM and CDMA can be combined to reduce the jamming effect. CDMA has anti-jamming capability which can be used with OFDM. This is very new technique and implementation is possible.

8 Conclusion

The performance of the WiMAX (IEEE 802.16) is degraded under jamming. Jamming can be detected easily but it affects the system badly and sometimes system may fail. In this paper, the performance of IEEE802.16-2004 based system under both multi-tone pilot and partial-band jamming are evaluated with the aid of computer simulations. The results show that the simulated throughput versus jamming power curves descends faster for weaker modulation/coding modes, indicating that weaker modes possess lower tolerance to jamming. Saturation phenomena to most of the jamming scenarios are observed for system operating. It proves the necessity and robustness of this newly defined operation mode. Multi-tone pilot jamming affects the system more severely than partial-band jamming under the same jamming power. In multi-tone pilot jamming, the number of jammed pilots decides the jamming severity and jamming 8 (full) pilots degrades the system to the most. In partial band jamming, system performance is degraded by either increased power of sub jammer or increased number of sub jammer, depending on the level of jamming power.

References

1. Andrews, J.G., Ghosh, A., Muhamed, R.: Fundamentals of WiMAX Understanding Broadband Wireless Networking. Prentice-Hall, Englewood Cliffs
2. Haykin, S., Moher, M.: Modern Wireless communication. Prentice-Hall, Englewood Cliffs

3. Ahson, S., Ilyas, M., Ahson, S., Ilyas, M.: WiMAX Standards and Security. CRC Press, Boca Raton (2008)
4. Shon, T., Choi, W.: An Analysis of Mobile WiMAX security: Vulnerabilities and Solutions (2007)
5. Cuilan, L.: A Simple Encryption Scheme Based on WiMAX, Department of Electronics Jiangxi University of Finance and Economics Nanchang, China (2009)
6. Makarevitch, B.: Jamming Resistant Architecture for WiMAX Mesh Network, communications Laboratory Helsinki University of Technology.
7. Li, J., Häggman, S.-G.: Performance of IEEE 802.16 Based System in Jamming Environment and Its Improvement With Link Adaption. In: The 17th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC 2006 (2006)
8. Nasreldin, M., Aslan, H., El-Hennawy, M., El-Hennawy, A.: WiMAX Security. In: 22nd International Conference on Advanced Information Networking and Applications (2008)
9. White paper by Motorola, WiMAX Security for Real- World Network Service Provider Deployments (2007)
10. IEEE 802.16 Working Group. IEEE 802.16-2004 Local and metropolitan area networks - Part 16: Air interface for fixed broadband wireless access systems IEEE Standard for Local and Metropolitan Area Networks. IEEE Computer Society Press, Los Alamitos
11. IEEE 802.16 Working Group. IEEE 802.16e-2005 IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment for Physical and Medium Access Control Layers

CE-OFDM: A PAPR Reduction Technique

R.S. Chaudhari¹ and A.M. Deshmukh²

¹ M.E. Student, Sinhgad College of Engineering, Pune – 41, Maharashtra, India

² Associate Professor, Sinhgad College of Engineering, Pune – 41, Maharashtra, India

rupali.chaudhari06@gmail.com

achala.deshmukh@gmail.com

Abstract. Orthogonal Frequency Division Multiplexing (OFDM) is a multi-carrier modulation technique having multiple subcarriers which can be overlapped. It finds various applications in wireless communication networks. The main advantages of OFDM are in high rate data communications; but the major hurdle is that it suffers from the Peak-to-Average Power Ratio (PAPR) problem. These causes inter modulation distortion as well as out-of-band spectral growth. Constant envelope OFDM (CE-OFDM) transforms the OFDM signal, to a signal with nearly *zero* dB PAPR. The performance of conventional OFDM system improves with constant envelope OFDM system.

1 Introduction

To provide broadband wireless communication network systems there should be a provision to meet the huge rise in the demand for flexible high data-rate services. These high data-rates with a very high signal quality depend a lot on the wireless channels. But there is always a scarcity of the channel bandwidth available. Again it is difficult to utilize the bandwidth effectively unless efficient techniques are employed. Furthermore, the wireless channels have impairments such as fading, shadowing and multi-user interference which highly degrade the system performance.

OFDM is a parallel-data-transmission scheme, which reduces the influence of multipath fading]. OFDM is robust against narrowband interference because such interference affects only a small percentage of the subcarriers; also it increases robustness against frequency-selective fading.

But the primary shortcoming of OFDM is that the modulated waveform has high amplitude fluctuations that produce large peak-to-average power ratios (PAPRs). This high PAPR makes OFDM sensitive to nonlinear distortion caused by the transmitter's power amplifier. Without sufficient power backoff, the system suffers from spectral broadening, intermodulation distortion, and, consequently, performance degradation. These problems can be reduced by increasing the backoff, but this result in reduced power amplifier efficiency [6, 12].

Many techniques have been developed to address the PAPR problem. Different techniques provide different degrees of effectiveness, and present different sets of tradeoffs. An alternative approach to mitigate this PAPR problem is based on signal transformation. Constant envelope OFDM (CE-OFDM) is a technique that transforms the OFDM signal with phase modulation, to a signal designed for efficient power

amplification. The constant envelope OFDM system shares many of the same functional blocks as the conventional OFDM system.

Section 2 explains the advantages of OFDM and its primary disadvantage of high PAPR. The different techniques available for reducing this PAPR are also specified. It is followed by the concept of constant envelope. The transmitter and the receiver of the constant envelope system are explained. Section 3 includes the response of the system and the performance of the CE-OFDM system as compared to the conventional OFDM system. Section 4 is the conclusions followed by the references.

2 Constant Envelope OFDM System

An OFDM signal consists of a number of independently modulated subcarriers. All subcarriers are placed orthogonally with respect to each other. This is the main advantage of OFDM that it utilizes the bandwidth effectively and efficiently. When N such subcarriers are added up coherently with the same phase, they produce a peak power that is N times the average power.

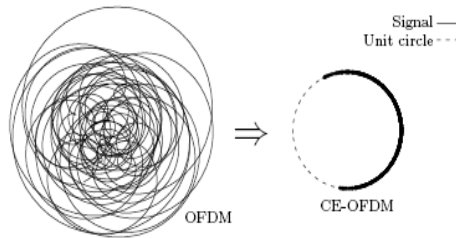


Fig. 1. CE-OFDM waveform mapping

A large peak-to-average power ratio brings disadvantage like an increased complexity of the transmitter and receiver. It also reduces the efficiency of the system. To reduce the PAP ratio, different techniques exist. There are signal distortion techniques, which reduce the peak amplitudes simply by nonlinearly distorting the OFDM signal at or around the peaks. Examples are clipping, peak windowing [4].

Secondly, there are coding techniques that use a special FEC code set that excludes OFDM symbols with a large PAP ratio. Examples are block coding schemes, block coding scheme with error correction [1]. The third technique, scrambling, scrambles each OFDM symbol with different scrambling sequences; selecting the sequence that gives the smallest PAP ratio. Examples include selected mapping [2], partial transmit sequence [3].

Here, in CE-OFDM this reduction in PAPR is achieved through signal transformation technique. This technique includes phase modulation at the transmitter and phase demodulation at the receiver. Phase modulation transforms the amplitude variations into a constant amplitude signal thereby reducing the vast difference between the peak power and average power.

The Constant envelope OFDM (CE-OFDM) is mapping of the OFDM signal to a unit circle [11], as depicted in Fig. 1. The instantaneous power of the resulting signal is constant. For the CE-OFDM signal the peak and average powers are the same, thus the PAPR is 0 dB. The transmitter and the receiver of the CE-OFDM are explained below.

2.1 Transmitter Side

At the transmitter side as shown in the Fig. 2, the input data is in the form of source bits. The input data is firstly mapped and then are converted from a serial stream to parallel sets. Each set of data contains one symbol, S_i , for each subcarrier.

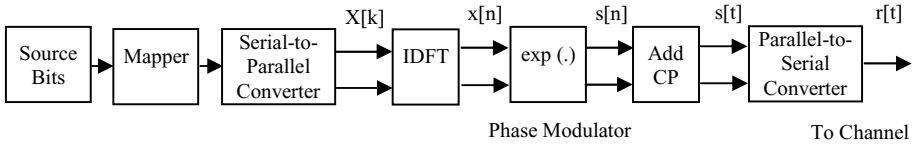


Fig. 2. CE-OFDM System Model - Transmitter

During each T -second block interval, an N -DFT point Inverse Discrete Fourier Transform (IDFT) calculates a block of time samples $\{x[n]\}$. Then, $\{x[n]\}$ which is a high PAPR OFDM sequence is passed through a phase modulator to obtain the 0 dB PAPR sequence [12]. In the above block diagram, the $\exp(\cdot)$ block acts as a phase modulator which actually does PAPR reduction.

Further the cyclic prefix (CP) i.e. appended to $\{s[n]\}$ to obtain the actual signal to be transmitted. Then, the parallel-to-serial converter creates the OFDM signal by sequentially outputting the time domain samples. The discrete-time samples are then transmitted into the channel. The channel impulse response is modeled as a wide-sense stationary uncorrelated scattering (WSSUS) process [9] comprised of L discrete paths [5, 11].

2.2 Receiver Side

At the receiver end as shown in the Fig. 3, the cyclic prefix samples are discarded firstly and the remaining samples $\{r[n]\}$ are processed [12]. The OFDM data is converted from a serial stream into parallel sets. A frequency-domain equalizer (FDE) is used to correct the distortion caused by the channel. Once the received signal has been equalized in the frequency domain, each received signal frequency component is multiplied by a complex coefficient (tap) $C[k]$ where the taps are obtained from the channel frequency response [13]. The final step entails transforming it back to the time domain, $\{s^{\wedge}[n]\}$. This is accomplished through application of the IDFT. Then the parallel set of data is converted to a serial stream of data.

At the phase demodulator, the phase is extracted by taking the inverse tangent [9] of the quadrature baseband components using the $\arctan(\cdot)$ block. This is followed by a phase unwrapper to obtain the demodulated phase $\{x^{\wedge}[n]\}$ through the unwrap block.

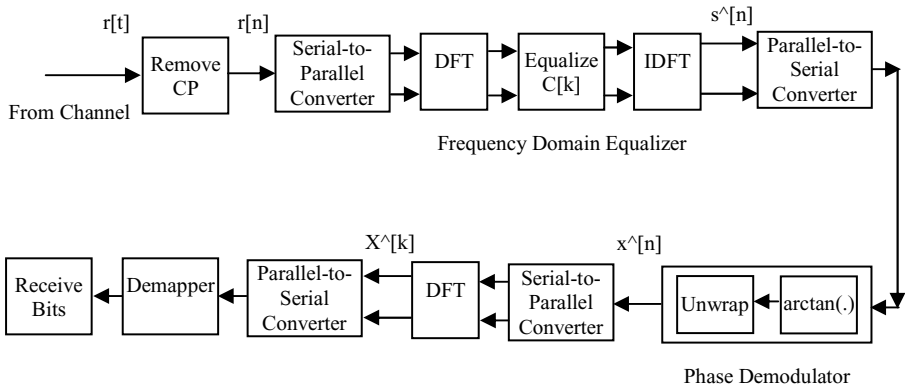


Fig. 3. CE-OFDM System Model – Receiver

The output of the phase demodulator is in serial form which is converted to parallel set. Then it is processed by the OFDM demodulator which consists of the N correlators, one corresponding to each subcarrier. This correlator bank is implemented in practice with the Fast Fourier Transform resulting in $\{X^{\wedge}[k]\}$, and then the parallel-to-serial block converts this parallel data into a serial stream followed by a symbol demapper that yields the received bits.

Thus, the transmitter achieves reduction in PAPR by modulating the phase, to nearly 0 dB after conventional OFDM modulation. At the receiver side, phase demodulation is done first followed by the conventional OFDM demodulation. Thus, the signal transformation technique works.

3 CE-OFDM System Response

The Baseband OFDM waveform can be represented by equation (1)

$$v(t) = A_v \sum_{k=0}^{N-1} d_k e^{\frac{j2\pi kt}{T}} \tag{1}$$

for $0 \leq t \leq T$ where A_v , is a gain constant, N is the number of subcarriers, T is the signaling interval, and data symbol d_k which modulates the k^{th} subcarrier $e^{\frac{j2\pi kt}{T}}$ [10].

The data symbols are chosen from a complex set defined by an M -point signal constellation such as PSK or QAM. Random phase alignment of the subcarriers results in large signal peaks. OFDM subcarriers when summed together show tremendous variations in the amplitude giving rise to peak power and average power problem.

Let us see the performance of the CE-OFDM phase demodulator. The performance of the demodulator improves in CE-OFDM as compared to conventional OFDM system. The phase demodulator consists of an arctangent calculator and a phase unwrapper. The performance of the receiver improves if a finite impulse response filter is

placed before the arctangent calculator [7, 11]. The performance is also dependent on the modulation index. The Symbol error rate is given by equation (2)

$$SER \approx 2 \left(\frac{M-1}{M} \right) Q \left(2\pi h \sqrt{\frac{6 \log_2 M}{M^2 - 1}} \frac{E_b}{N_o} \right) \tag{2}$$

where Q is the Gaussian Q -function, E_b is the energy per bit of the transmitted band-pass CE-OFDM signal and N_o is the additive Gaussian noise [8]. And the Bit Error Rate is given by equation (3)

$$BER \approx \frac{SER}{\log_2 M} \tag{3}$$

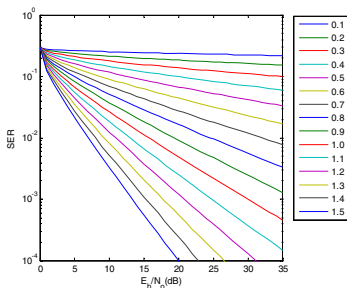


Fig. 4. Performance of the CE-OFDM Phase Demodulator Receiver ($M = 8$)

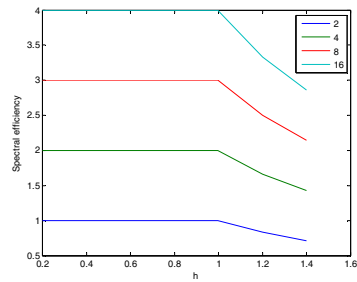


Fig. 5. Spectral Efficiency for Modulation Indices ($M = 2, 4, 8, 16$)

The Fig. 4 shows BER Vs Eb/No for different modulation indices. It shows BER decreases with increase in the modulation index. It depicts that when the modulation index is small, the bit energy-to-noise ratio is high. For larger modulation index, the analytical expression becomes less accurate and for modulation index, $h = 1.5$, an irreducible error floor develops.

This illustrates the limitation of discrete-time arctangent phase demodulator. Properly phase unwrapping a noisy signal is a difficult problem. Occasional phase jumps at the output of the phase demodulator corrupts the demodulated phase angle and degrades BER performance.

The performance of CE-OFDM system largely depends on the modulation index, h of the modulator. The spectral efficiency, S improves with modulation index but after certain value of modulation index the efficiency decreases [12]. The spectral efficiency can be calculated by the equation (4)

$$S = \frac{R}{B} = \frac{\log_2 M}{\max(2\pi h, 1)} \tag{4}$$

where M , is the modulation order.

The Fig. 5 is a plot of Spectral efficiency, S against Modulation index, h for different values of Modulation order, M . It can be seen that for fixed values of modulation

index, the CE-OFDM system has improved spectral efficiency with increased modulation order, M ; but at the cost of performance degradation.

Performance analysis is extended to the case of fading channels. It is desired to calculate the bit error rate at a given average bit energy-to-noise density ratio [12]. This quantity depends on the statistical distribution of bit energy-to-noise density ratio, γ . For channels without a line-of-sight (LOS) component, the power of LOS tends to zero and bit energy-to-noise density ratio, γ becomes Rayleigh distributed.

At high CNR, the earlier BER equation provides an accurate expression for the conditional BER. An approximate lower bound on BER is therefore obtained by assuming that earlier BER equation is accurate for all CNR. The Fig. 6 is a plot of BER against Bit Energy-to-Noise Density ratio for the channel, γ for BER for lower bound and BER for upper bound. It can be seen that as the modulation index, h increases the BER decreases. But still the BER is for more high than expected because the channel considered is non-frequency selective. The performance improves with frequency-selective channels.

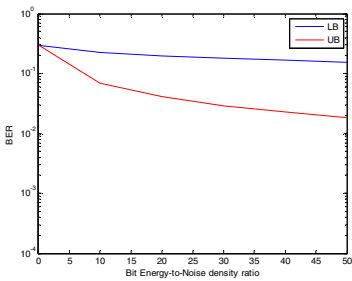


Fig. 6. Performance of CE-OFDM in Flat Fading channels

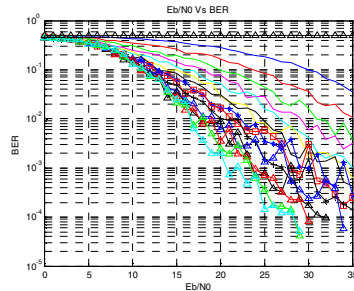


Fig. 7. Comparison of Conventional OFDM with CE-OFDM ($M = 8, N = 64, SNR_{max} = 35$)

The performance of any system can be better judged with its Bit error rate. Finally the conventional and CE-OFDM systems are compared in terms of Bit error rate for different values of Bit energy to noise density ratio. The performance of CE-OFDM system improves because of PAPR reduction by means of phase modulation [12].

The CE-OFDM system is simulated; Fig. 7 is a plot of BER Vs E_b/N_0 for different values of Modulation index. It compares the performance of CE-OFDM and conventional OFDM system. It shows that with CE-OFDM the BER reduces. The topmost horizontal line shows the performance of the conventional OFDM system.

The bit error rate is very high for conventional OFDM system because the system suffers from intermodulation distortion and high amplitude variations which further degrades the performance of the system. These shortcomings are very well overcome in the constant envelope OFDM system.

4 Conclusions

A transformation technique that eliminates the PAPR problem associated with OFDM is considered. The phase modulation transforms the high peak-to-average ratio to 0 dB PAPR. PAPR reduction occurs without any severe distortions as in other techniques. CE-OFDM is studied over a wide range of channel conditions. For frequency-selective fading channels, a frequency-domain equalizer improves the performance of the receiver. CE-OFDM compares favorably to conventional OFDM in multipath fading channels. However CE-OFDM suffers from the FM threshold effect. The above problem can be alleviated with threshold extension techniques. Most notably, phase locked loops are known to reduce the threshold in FM systems by several dB. Realizing such an enhancement, and determining its complexity is very complex.

References

- [1] Ahn, H., Shin, Y.M., Im, S.: A block coding scheme for peak to average power ratio reduction in an orthogonal frequency division multiplexing system. In: IEEE Vehicular Technology Conference Proceedings, vol. 1 (2000)
- [2] Bauml, R.W., Fischer, R.F.H., Huber, J.B.: Reducing the peak to average power ratio of multi carrier modulation by selective mapping. IEEE Electronics Letters 32 (1996)
- [3] Muller, S.H., Huber, J.B.: OFDM with reduced peak to average power ratio by optimum combination of partial transmit sequences. IEEE Electronics Letters 33 (1997)
- [4] Van Nee, R., Wild, A.: Reducing the peak to average power ratio of OFDM. In: IEEE Vehicular Technology Conference Proceedings, vol. 3 (1998)
- [5] Patzold, M.: Mobile Fading Channels. John Wiley & Sons, West Sussex (2002)
- [6] Prasad, R.: OFDM for Wireless Communications Systems. Artech House, Inc., Boston (2004)
- [7] Proakis, J.G.: Digital Communications, 4th edn. McGraw - Hill, New York (2001)
- [8] Proakis, J.G., Salehi, M.: Communication Systems Engineering. Prentice Hall, New Jersey (1994)
- [9] Proakis, J.G., Manolakis, D.G.: Digital Signal Processing: Principles Algorithms, and Applications, 3rd edn. Prentice Hall, Upper Saddle River (1996)
- [10] Thompson, S.C., Proakis, J.G., Zeidler, J.R.: Constant Envelope Binary OFDM Phase Modulation. In: Proceedings of IEEE MILCOM, Boston (2003)
- [11] Thompson, S.C., Proakis, J.G., Zeidler, J.R.: Noncoherent reception of constant envelope OFDM in flat fading channels. Proceedings of IEEE PIMRC 1, 517–521 (2005)
- [12] Thompson, S.C., Ahmed, A.U., Proakis, J.G., Zeidler, J.R., Geile, M.J.: Constant Envelope OFDM. IEEE Transactions on Communications 56(8) (2008)
- [13] Thompson, S.C., Proakis, J.G., Zeidler, J.R.: Channel estimation and equalization for CE-OFDM in multipath fading channels. IEEE Transaction (2008), 978-1-4244-2677-5/08

Initializing Cluster Center for K-Means Using Biogeography Based Optimization

Vijay Kumar¹, Jitender Kumar Chhabra², and Dinesh Kumar³

¹ CSE Department, JCDMCOE, Sirsa, Haryana, India

² Computer Engg. Department, NIT, Kurukshetra, Haryana, India

³ CSE Department, GJUS&T, Hisar, Haryana, India

vijaykumarchahar@gmail.com, jitenderchhabra@gmail.com,
dinesh_chutani@yahoo.com

Abstract. In spite of K-Means algorithm of clustering problems being widely used, it suffers from problem of initialization of centroids being done randomly. In this paper, we propose a hybrid approach to solve the problem of random initialization of cluster centers or centroids using BBO – a population based evolutionary algorithm motivated by migration mechanism of ecosystems. The paper further determines the optimal number of clusters using different cluster validity indexes such as Rand, Mirkin, Hubert, Jaccard and FM indexes. Experimental results demonstrate that proposed approach outperforms the K-Means, K-Mediod and Gustafson- Kessel clustering algorithms in terms of precision, recall and G-measure.

Keywords: Clustering, BBO, K-Means.

1 Introduction

Clustering is the process of grouping together the similar data into a number of clusters. These techniques have been applied in different fields of science and engineering such as data mining, image segmentation and pattern recognition [1, 2]. These can be roughly classified into two categories: hierarchical clustering and partition clustering. Hierarchical clustering techniques are able to find structures which can be further divided in substructures and so on recursively [1], some of which are: Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), Clustering using REpresentatives (CURE) [5], CHEMELON [6] and ROCK [7]. Partition clustering techniques try to obtain a single partition of data without any other sub-partition and are based on the optimization of an objective function [23]. The most popular partition clustering techniques are: K-Means [4], K-Mediod [22], Fuzzy C-Means, SOM and Expectation-Maximization. Among these, K-Means technique has gained more popularity due to its simplicity and efficiency, but it suffers from being sensitivity to initialization of cluster center and also has weakness of converging to local minima. To solve these problems, many heuristic clustering techniques have been developed.

In this paper, an implementation of BBO to improve the K-Means has been done. BBO is used to initialize the cluster center for K-Means. After that, K-Means algorithm is applied to data. The rest of the paper is organized as follows. Section 2 presents the

general K-Means algorithm. Section 3 gives brief description of BBO. Section 4 presents proposed initial cluster refinement procedure. Section 5 presents the quality measurement metrics. Section 6 covers the experimental results followed by conclusions in section 7.

2 K-Means Algorithm

The important component of any clustering algorithm is the measure of similarity. In K-Means, Euclidean distance is used as similarity measure for clustering the data set into a predefined number of clusters [13]. Data points within cluster having smallest euclidean distance are compared to neighboring clusters.

K-Means divides a given data set into K clusters through iterative procedure. The procedure consists of following steps [4]:

1. Initialize K centroids $(c_i, 1 \leq i \leq K)$ in the vector space.
2. Calculate the distances from every point to every centroid. Assign each point to group i , if c_i is its closet centroid.
3. Update centroids. Each centroid is updated as the mean of all the points in its group.
4. If no point changed its membership, exit, otherwise, go to step 3.

One of the major problems in K-Means is that it is sensitive to the cluster center points selected initially, leading to production of different results based upon different values of initialization [3]. The results also depend upon the number of clusters, which is mostly not known or determined in advance.

There is no existing general theoretical solution to find optimal number of clusters for any dataset [3]. A simple approach is to rank the results of multiple runs using different values of cluster number.

3 Biogeography Based Optimization

The concept of Biogeography Based Optimization (BBO) was first presented by D. Simon in 2008. It is a population based evolutionary algorithm that is based on the mathematics of biogeography [8, 10]. Biogeography is the study of geographical distribution of biological organisms. Biogeography describes how species migrate from one island/habitat to another, how new species arise, and how species become extinct [8]. Geographical areas that are most suitable for biological species possess high habitat suitability index (HSI). Features, that are correlated with HSI, include food availability, drought situation etc. The variables which characterize habitability are called suitability index variables (SIV) [8, 9]. High HSI habitats are more static than the low HSI habitats. A good solution is analogous to habitats having high HSI. Similarly, habitats having low HSI value are considered poor solutions.

In BBO, the problem and a population of candidate solutions are represented by vector of integers [8]. Each integer in solution vector is considered as a SIV. Those solutions

are considered good that have higher HSI values. Low HSI accepts a lot of new features from high HSI [10]. Emigration and migration rate of each solution is used to share information between habitats using probability of modification. Each solution can be modified based on other solutions. If given solution is selected to be modified, then immigration rate decides whether or not to modify each SIV in that solution [8]. If SIV is selected to modify, then we use emigrate rate of other solutions to decide which of the solutions should migrate a selected SIV to given solution [8, 10].

The BBO algorithm consists of following steps [8]:

1. Initialize the maximum species count, maximum emigration rate, maximum immigration rate, mutation rate and elitism parameter.
2. Initialize a random set of habitats; each habitat corresponding to solution to the given problem.
3. For each habitat, map the HSI to the number of species, immigration rate and emigration rate.
4. Probabilistically use immigration and emigration to modify each non-elite habitat, then recomputed each HSI.
5. For each habitat, update the probability of its species count. Then mutate each non-elite habitat based on its probability and recomputed each HSI.
6. Go to step 3 for the next iteration. This loop can be terminated after a predefined number of generations or after an acceptable problem solution has been found.

4 Proposed Initial Cluster Refinement Procedure

The proposed cluster refinement procedure initializes K-Means using Biogeography Based Optimization (BBOKMI). In BBOKMI, BBO is executed on dataset to provide the cluster centers. These are used to initialize cluster centers for K-means algorithm. In BBO, each chromosome is represented by the centers of K clusters. For N data-points of M dimensional space, the length of a Chromosome will be $M \times K$ bits/position, where the first M positions correspond to the center of first cluster, the next M positions correspond to those of the centre of second cluster, and so on. The K clusters encoded in each chromosome are initialized. The selection of chromosomes has been done randomly from datasets. This procedure is repeated as many times as the size of the population. HSI are assigned values equal to smallest euclidean distance which calculated between the chromosome of population and all data points. BBO was applied on whole population to get the optimal value of HSI, which can be later used to initialize the K-Means algorithm.

5 Quality Measurement Matrices

5.1 Cluster Validity Analysis

We use six different validity indexes for finding the optimal number of clusters. These validity indexes are: Rand Index [14], Jaccard Index [15], Hubert Index [16],

Adjusted Rand (AR) Index [17], Folowes and Mallows (FM) Index [18] and Mirkin Index [19]. Rand, Jaccard, Hubert, Adjusted Rand and Fowlkes and Mallows indexes give largest value for optimal number of clusters whereas Mirkin index gives smallest value, when number of clusters attains optimal value.

5.2 Cluster Quality Analysis

Weighted average, precision, recall and G-Measure [20, 21] have been used for goodness/quality of clustering. Precision is a measure of the ability of system to present only relevant items. Recall is a measure of the ability of the system to present all the relevant items. Weighted average is the ratio of correctly claimed classes to the total number of classes.

6 Experimentation and Results

6.1 Datasets Used

All the clustering techniques used in this paper have been applied on two real-life datasets of UCI database [24]. The real life datasets are: “Iris” and “Wine” datasets. Table 1 presents the details of these datasets.

Table 1. UCI datasets

Dataset	No. of data points	No. of features	No. of Classes
Iris	150	4	3
Wine	178	13	3

6.2 Parameters Used

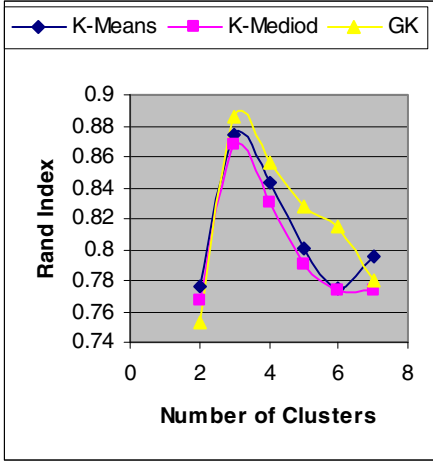
For Gustafson Kessel (GK) clustering, we set weighting exponent and tolerance error to 2 and 0.0001 respectively. For BBOKMI, we keep population size as 20 and mutation probability is 0.023. Emigration and immigration rates for BBOKMI are 1. Lower bound and upper bound for immigration rate are 0 and 1 respectively.

6.3 Experiment 1 and Results

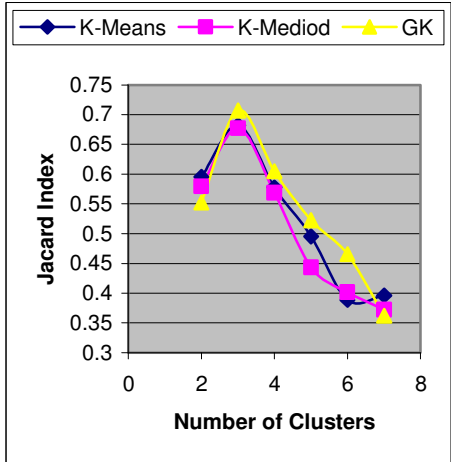
First, we calculate the optimal number of clusters required for iris and wine datasets. For this, we use six validity indexes for K-Means, K-Mediod and GK [11, 12] clustering techniques. Figures 1(a-e) show the effect of varying number of clusters on validity indexes for iris dataset. The results reveal that the best values for all cluster validity indexes are achieved when number of cluster is 3 for all clustering techniques. Figures 2 (a-e) show validity indexes for wine dataset.

From figures 1(a) and 1(b), Rand and Jaccard indexes attain largest value when the number of clusters is 3 for all above said methods. Figures 1(c), 1(d) and 1(e) also show that Hubert, Adjusted Rand and FM Indexes are largest for 3 as the number of

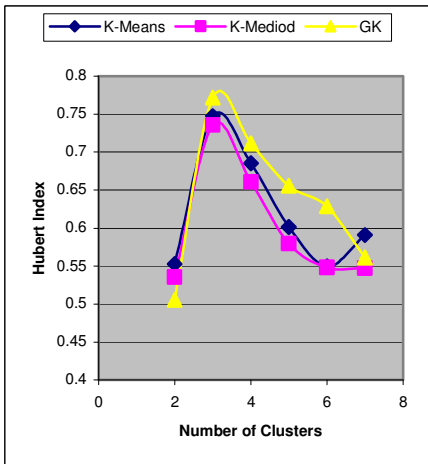
clusters. All indexes give largest value at number of clusters is equal to 3. Mirkin index gives optimal number of clusters when index attains minimum value. We observed from figure 1(f) that Mirkin index having lowest value at number of clusters is 3. So, the optimal number of clusters for all above said techniques is 3 for iris dataset.



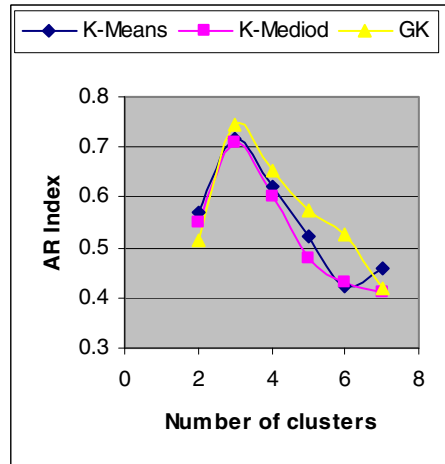
(a)



(b)

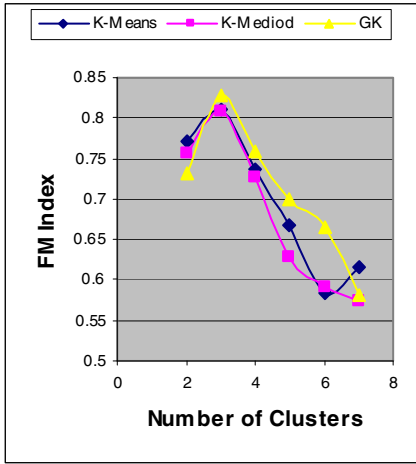


(c)

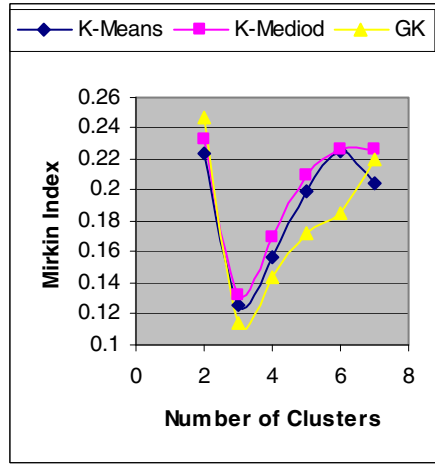


(d)

Fig. 1. Cluster Validity Indices for Iris dataset; (a) Rand (b)Jaccard (c) Hubert (d) Adjusted Rand (e) Fowlkes and Mallows (f) Mirkin



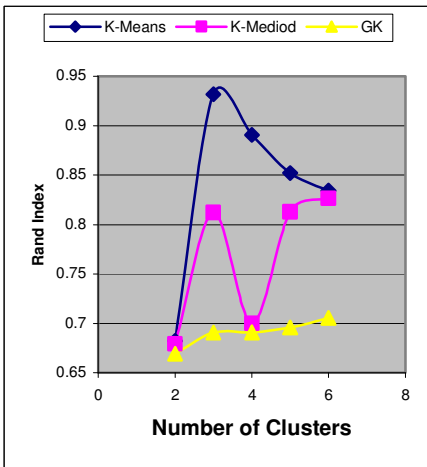
(e)



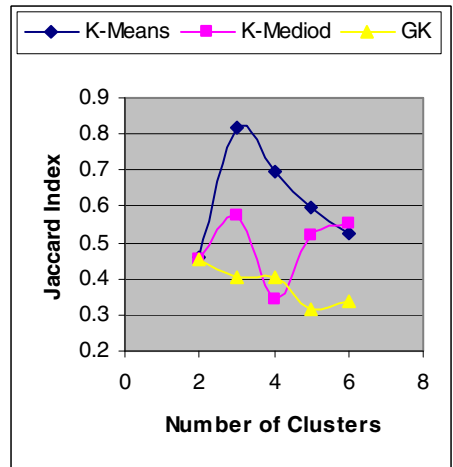
(f)

Fig. 1. (Continued)

The results from figures 2(a-f) show that optimal number of clusters required for wine dataset is 3 for K-Means and K-Mediod, but GK does not show constant results over validity indexes.

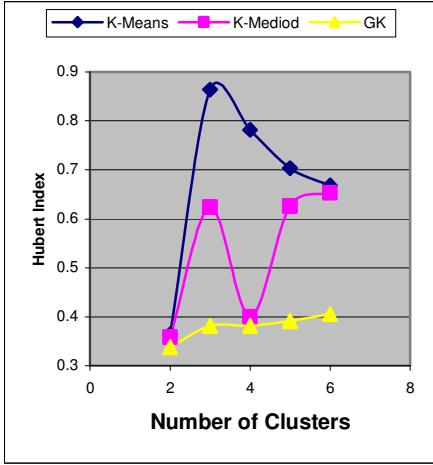


(a)

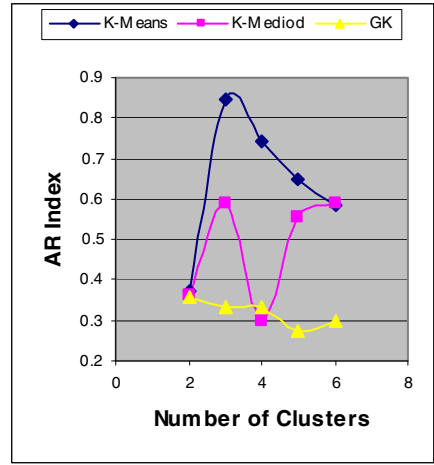


(b)

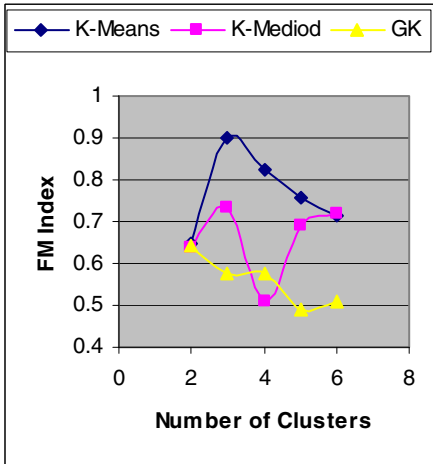
Fig. 2. Validity Indices for Wine dataset; (a) Rand (b)Jaccard (c) Hubert (d) Adjusted Rand (e) Fowlkes and Mallows (f) Mirkin



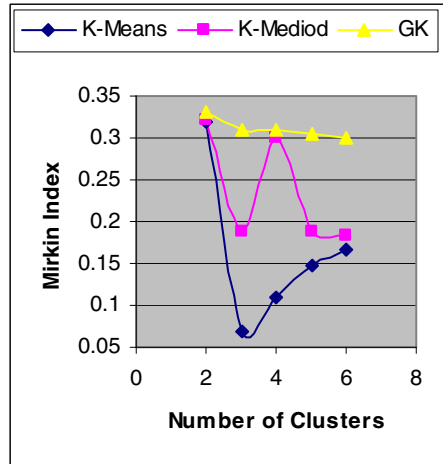
(c)



(d)



(e)



(f)

Fig. 2. (Continued)

6.4 Experiment 2 and Results

To evaluate the performance of BBOKMI technique, we compared it with K-Means (KM), K-Mediod (KMD) and GK Clustering techniques. These were applied on both datasets. These datasets also have a class label for classification purpose. We use optimal number of clusters (i.e. 3) for each dataset. Table 2 shows the comparison between proposed BBOKMI approach and above said techniques in terms of weighted average. The values of G-Measure, precision and recall are tabulated in tables 3, 4 and 5. The results reveal that BBOKMI outperforms K-Means, K-Mediod and GK clustering in terms of precision, recall, weighted average and G-measure. In wine dataset, the results of recall are better for GK clustering.

Table 2. Weighted Average for different clustering algorithms

	K-Means	K-Mediod	GK	BOKMI
Iris	0.0200	0.2733	0.3333	0.88667
Wine	0.3483	0.1292	0.6067	0.70225

Table 3. G-Measure for different clustering algorithms

	K-Means	K-Mediod	GK	BOKMI
Iris	0	0	0	0.88527
Wine	0	0.0838	0.5660	0.69225

Table 4. Precision for different clustering algorithms

	K-Means	K-Mediod	GK	BOKMI
Iris	0.0256	0.2733	0.1853	0.8980
Wine	0.3330	0.1236	0.6100	0.7230

Table 5. Recall for different clustering algorithms

	K-Means	K-Mediod	GK	BOKMI
Iris	0.020	0.3176	0.6667	0.8867
Wine	0.291	0.1196	0.7993	0.6955

7 Conclusion

In this paper, we have proposed a BBO based initialization of the K-Means Clustering. The optimal number of clusters has been calculated for two datasets using different validity indexes such as Rand, Jaccard, Adjusted Rand, Mirkin, Hubert and Fowlkes and Mallows indexes. Experimental results demonstrated that the optimal number of clusters is three for iris and wine datasets. On comparing the results of proposed technique with the existing ones, it has been found that BBOIKM performs better than K-Means, K-Mediod and Gustafson-Kessel Clustering techniques.

References

1. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A review. *Int. J. ACM Computing Surveys* 31, 264–323 (1999)
2. Mary, C.I., Raja, S.V.K.: Refinement of clusters from K-Means with Ant Colony Optimization. *J. Theor. and Applied Info. Tech.* 10, 28–32 (2009)
3. Al-Shboul, B., Myaeng, S.H.: Initializing K-Means using Genetic Algorithm. *J. WASET* 54, 114–118 (2009)
4. Lei, H., Tang, L.P., Iglesias, J.R., Mukherjee, S., Mohanty, S.: S-Means: Similarity Driven Clustering and its application in Gravitational-Wave Astronomy Data Mining. In: *Int. Workshop on Knowledge Discovery from Ubiquitous Data Streams*, pp. 25–36 (2008)

5. Guha, S., Rastogi, R., Shim, K.: CURE: An Efficient Clustering Algorithm for Large Data Sets. In: ACM SIGMOD Conference (1998)
6. Karypis, G., Han, E.H., Kumar, V.: CHAMELEON: a hierarchical clustering algorithm using dynamic modeling. *J. Computers* 32, 68–75 (1999)
7. Guha, S., Rastogi, R., Shim, K.: ROCK: A Robust Clustering Algorithm for Categorical Attributes. In: IEEE Conf. Data Engg. (1999)
8. Simon, D.: Biogeography-based Optimization. *IEEE Trans. Evolution. Comp.* 12(6), 702–713 (2008)
9. Wesche, T., Goertler, G., Hubert, W.: Modified habitat suitability index model for brown trout in southeastern wyoming. *North Amer. J. Fisheries Manage.* 7, 232–237 (1987)
10. Simon, D.: A Probabilistic Analysis of a Simplified Biogeography-Based Optimization Algorithm. *IEEE Trans. Evolution. Comp.* (2009)
11. Gustafson, D.E., Kessel, W.C.: Fuzzy Clustering with fuzzy covariance matrix. In: IEEE CDC, San Diego, pp. 761–766 (1979)
12. Babuska, R., Van der veen, P.J., Kaymak, U.: Improved covariance estimation for Gustafson Kessel Clustering. In: IEEE Conf. on Fuzzy Systems, pp. 1081–1085 (2002)
13. Van der merwe, D.W., Engelbrecht, A.P.: Data Clustering using Particle Swarm Optimization, pp. 215–220 (2003)
14. Rand, W.M.: Objective Criterion for Evolution of Clustering Methods. *J. American Stat. Assoc.* 66, 846–850 (1971)
15. Jaccard, P.: The distribution of flora in the alpine zone. *J. New Phytologist* 11, 37–50 (1912)
16. Hubert, L.: Kappa revisited. *J. Psych. Bulletin* 84, 289–297 (1977)
17. Hubert, L., Arabie, P.: Comparing partitions. *J. Classification* 2, 193–218 (1985)
18. Fowlkes, E.B., Mallows, C.L.: A Method for Comparing Two Hierarchical Clustering. *J. American Stat. Assoc.* 78, 553–569 (1983)
19. Mirkin, B.G., Cherny, L.B.: Deriving a distance between partitions of a finite set. *J. Auto. Remote Control* 31, 91–98 (1970)
20. Kowalski, G.: Information Retrieval Systems-Theory and Implementation. Kluwer Academic Publishers, Dordrecht (1997)
21. Buckland, M.K., Gey, F.: The relationship between Recall and Precision. *J. American Soc. Info. Sci.* 45, 12–19 (1994)
22. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press, USA (2006)
23. Filippone, M., Camastra, F., Masulli, F., Rovetta, S.: A survey of kernel and spectral methods for clustering. *J. Pattern Recog.* 41, 175–190 (2008)
24. Blake, C.L., Merz, C.J.: UCI Repository of Machine Learning (1998), <http://www.ics.uci.edu/~mllearn/databases/>

Implementation of Parallel Image Processing Using NVIDIA GPU Framework

Brijmohan Daga¹, Avinash Bhute², and Ashok Ghatol³

¹ Fr. Conceicao Rodrigues College of Engineering, Mumbai, India

² V.J.T.I. Mumbai, India

³ D. Y. Patil Group of Institution, Pune, India

bsdaga@yahoo.com, anbhute@gmail.com, ashok.ghatol@gmail.com

Abstract. We introduced a real time Image Processing technique using modern programmable Graphic Processing Units (GPU) in this paper. GPU is a SIMD (Single Instruction, Multiple Data) device that is inherently data-parallel. By utilizing NVIDIA's new GPU Programming framework, "Compute Unified Device Architecture" (CUDA) as a computational resource, we realize significant acceleration in the computations of different Image processing Algorithms. Here we present an efficient implementation of algorithms on the NVIDIA GPU. Specifically, we demonstrate the efficiency of our approach by a parallelization and optimization of the algorithm. In result we show time comparison between CPU and GPU implementation.

Keywords: GPU, CUDA, Image blending.

1 Introduction

Most powerful CPUs having multi-core processing power are not capable to attain Real-time image processing. Increasing resolution of video captures devices and increased requirement for accuracy make it harder to realize real-time performance. Recently, graphic processing units have evolved into an extremely powerful computational resource. For example, The NVIDIA GeForce GTX 280 is built on a 65nm process, with 240 processing cores running at 602 MHz, and 1GB of GDDR3 memory at 1.1GHz running through a 512-bit memory bus. Its Peak processing power is 933 GFLOPS [1], billions of floating-point operations per second, in other words. As a comparison, the quad-core 3GHz Intel Xeon CPU operates roughly 96 GFLOPS [2]. The annual computation growth rate of GPUs is approximately up to 2.3x. In contrast to this, that of CPUs is 1.4x [2]. At the same time, GPU is becoming cheaper and cheaper.

As a result, there is strong desire to use GPUs as alternative computational platforms for acceleration of computational intensive tasks beyond the domain of graphics applications. To support this trend of GPGPU (General-Purpose Computing on GPUs) computation [3], graphics card vendors have provided programmable GPUs and high-level languages to allow developers to generate GPU-based applications.

In this paper we demonstrate a GPU-based implementation of pyramidal blending algorithm implemented on NVIDIA's CUDA (Compute Unified Device Architecture). In Section 2, we describe the recent advances in GPU hardware and programming framework, we also discuss previous efforts on application acceleration using CUDA

framework, and the use of GPUs in computer vision applications. In Section 3, we detail the implementation of the pyramidal blending algorithm. In Section 4, we made various design and optimization choices for GPU-based Implementation of the algorithm, then we demonstrate the efficiency of our approach by applying it to CUDA framework.

2 Background

2.1 The NVIDIA CUDA Programming Framework

Traditionally, general-purpose GPU programming was accomplished by using a shader-based framework [4]. The shader-based framework has several disadvantages. This framework has a steep learning curve that requires in-depth knowledge of specific rendering pipelines and graphics programming. Algorithms have to be mapped into vertex transformations or pixel illuminations. Data have to be cast into texture maps and operated on like they are texture data. Because shader-based programming was originally intended for graphics processing, there is little programming support for control over data flow; and, unlike a CPU program, a shader-based program cannot have random memory access for writing data. There are limitations on the number of branches and loops a program can have. All of these limitations hindered the use of the GPU for general-purpose computing. NVIDIA released CUDA, a new GPU programming model, to assist developers in general-purpose computing in 2007 [3]. In the CUDA programming framework, the GPU is viewed as a compute device that is a co-processor to the CPU. The GPU has its own DRAM, referred to as device memory, and execute a very high number of threads in parallel. More precisely, data-parallel portions of an application are executed on the device as kernels which run in parallel on many threads.

In order to organize threads running in parallel on the GPU, CUDA organizes them into logical blocks. Each block is mapped onto a multiprocessor in the GPU. All the

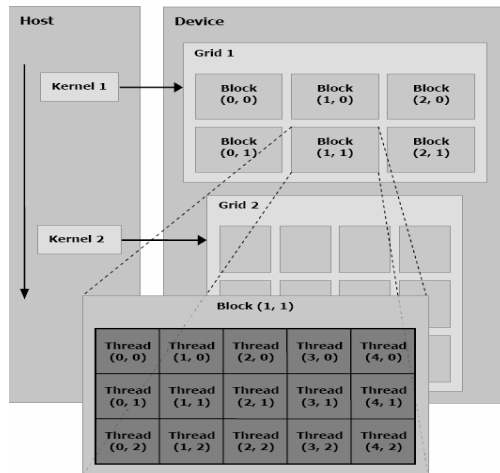


Fig. 1. Thread and Block Structure of CUDA

threads in one block can be synchronized together and communicate with each other. Because there are a limited number of threads that a block can contain, these blocks are further organized into grids allowing for a larger number of threads to run concurrently as illustrated in Figure 1. Threads in different blocks cannot be synchronized, nor can they communicate even if they are in the same grid. All the threads in the same grid run the same GPU code.

CUDA has several advantages over the shader-based model. Because CUDA is an extension of C, there is no longer a need to understand shader-based graphics APIs. This reduces the learning curve for most of C/C++ programmers. CUDA also supports the use of memory pointers, which enables random memory-read and write-access ability. In addition, the CUDA framework provides a controllable memory hierarchy which allows the program to access the cache (shared memory) between GPU processing cores a GPU global memory. As an example, the architecture of the GeForce 8 Series, the eighth generation of NVIDIA's graphics cards, based on CUDA is shown in Fig 2.

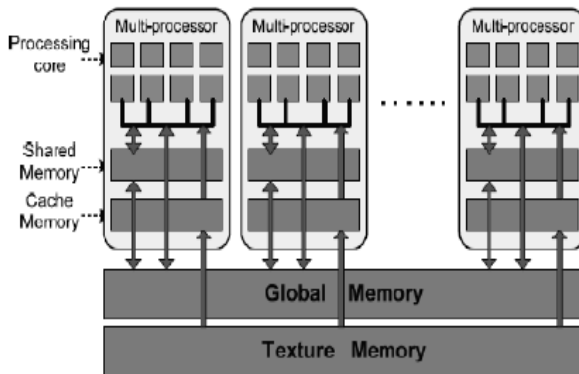


Fig. 2. GeForce 8 series GPU architecture

The GeForce 8 GPU is a collection of multiprocessors, each of which has 16 SIMD (Single Instruction, Multiple Data) processing cores. The SIMD processor architecture allows each processor in a multiprocessor to run the same instruction on different data, making it ideal for data-parallel computing. Each multiprocessor has a set of 32-bit registers per processors, 16KB of shared memory, 8KB of read-only constant cache, and 8KB of read-only texture cache. As depicted in Figure 2, shared memory and cache memory are on-chip. The global memory and texture memory that can be read from or written to by the CPU are also in the regions of device memory. The global and texture memory spaces are persistent across all the multiprocessors.

2.2 GPU Computation in Image Processing

Graphics Processing Units (GPUs) are high-performance many-core processors that can be used to accelerate a wide range of applications. Modern GPUs are very efficient at manipulating computer graphics, and their highly parallel structure makes them more effective than general-purpose CPUs for a range of complex algorithms. In a personal computer, a GPU can be present on a video card, or it can be on the motherboard. More

than 90% of new desktop and notebook computers have integrated GPUs, which are usually far less powerful than those on a dedicated video card. [1]

Most computer vision and image processing tasks perform the same computations on a number of pixels, which is a typical data-parallel operation. Thus, they can take advantage of SIMD architectures and be parallelized effectively on GPU. Several applications of GPU technology for vision have already been reported in the literature. De Neve et al. [5] implemented the inverse YCoCg-R colour transform by making use of pixel shader. To track a finger with a geometric template, Ohmer et al. constructed gradient vector field computation and canny edge extraction on a shader-based GPU which is capable of 30 frames per second performance. Sinha et al. [6] constructed a GPU-based Kanade-Lucas-Tomasi feature tracker maintaining 1000 tracked features on 800x600 pixel images about 40 ms on NVIDIA GPUs. Although all these applications show real-time performance at intensive image processing calculations, they do not scale well on newer generation of graphics hardware including NVIDIA' CUDA.

3 Pyramidal Blending

In Image Stitching application, once all of the input images are registered (align) with respect to each other, we need to decide how to produce the final stitched (mosaic) image. This involves selecting a final compositing surface (flat, cylindrical, spherical, etc.) and view (reference image). It also involves selecting which pixels contribute to the final composite and how to optimally blend these pixels to minimize visible seams, blur, and ghosting.

In this section we describe an attractive solution to this problem was developed by Burt and Adelson [7]. Instead of using a single transition width, a frequency adaptive width is used by creating a band-pass (Laplacian) pyramid and making the transition widths a function of the pyramid level. First, each warped image is converted into a band-pass (Laplacian) pyramid. Next, the masks associated with each source image are converted into a low pass (Gaussian) pyramid and used to perform a per-level feathered blend of the band-pass images. Finally, the composite image is reconstructed by interpolating and summing all of the pyramid levels (band-pass images).

3.1 Basic Pyramid Operations

Gaussian Pyramid: A sequence of low-pass filtered images G_0, G_1, \dots, G_N can be obtained by repeatedly convolving a small weighting function with an image [7,8]. With this technique, image sample density is also decreased with each iteration so that the bandwidth is reduced in uniform one-octave steps. Both sample density and resolution are decreased from level to level of the pyramid. For this reason, we shall call the local averaging process which generates each pyramid level from its predecessor a REDUCE operation. Again, let G_0 be the original image. Then for $0 < l < N$:

$G_l = \text{REDUCE} [G_{l-1}]$, which we mean,

$$G_l(I, j) = \sum_{m,n=1}^5 W(m, n) G_{l-1}(2i+m, 2j+n)$$

Laplacian Pyramid: The Gaussian pyramid is a set of low-pass filtered images. In order to obtain the band-pass images required for the multi resolution blend we subtract each level of the pyramid from the next lowest level. Because these arrays differ in sample density, it is necessary to interpolate new samples between those of a given array before it is subtracted from the next lowest array. Interpolation can be achieved by reversing the REDUCE process. We shall call this an EXPAND operation. Let G_l, k be the image obtained by expanding G_l , k times. Then

$$G_{l,0} = G_l \text{ which we mean,}$$

$$G_{l,k}(I, j) = 4 \sum_{m,n=-2}^2 G_{l,k-1}(2i+m/2, 2j+n/2)$$

Here, only terms for which $(2i + m)/2$ and $(2j + n)/2$ are integers contribute to the sum. Note that $G_{l,1}$ is the same size as G_{l-1} , and that $G_{l,l}$ is the same size as the original image. We now define a sequence of band-pass images L_0, L_1, \dots, L_N . For $0 < l < N$, $L_l = G_l - \text{EXPAND}[G_{l+1}] = G_l - G_{l+1}$. Because there is no higher level array to subtract from G_N , we define $L_N = G_N$. Just as the value of each node in the Gaussian pyramid could have been obtained directly by convolving the weighting function W_l with the image, each node of L_l can be obtained directly by convolving $W_l - W_{l+1}$ with the image. This difference of Gaussian-like functions resembles the Laplacian operators commonly used in the image processing, so we refer to the sequence L_0, L_1, \dots, L_N as the Laplacian pyramid.

3.2 Algorithm

- Step 1: Build Laplacian pyramids LA and LB from images A and B .
- Step 2: Build a Gaussian pyramid GR from selected region R .
- Step 3: Form a combined pyramid LS from LA and LB using nodes of GR as Weights. $LS(i, j) = GR(i, j) * LA(i, j) + (1 - GR(i, j)) * LB(i, j)$
- Step 4: Collapse (by expanding and summing) the LS pyramid to get the final blended Image.

4 Proposed Implementation Details

In this section we describe various implementation strategies of the algorithm. We need to find possible parallelization in different functions of the algorithm. Pyramidal blending requires construction of Gaussian and Laplacian pyramid which are following the SIMD paradigm.

We set the execution configuration depending on size of shared memory of CUDA Memory hierarchy as it is the essential to execute Threads parallel. Number of blocks each multiprocessor can process depends on how many registers per thread and how much shared memory per block is required for a given kernel. Since shared memory is not used in the implementation with texture memory, we only need to be concerned about the number of registers used and we can maximize the size of block and grid as much as possible.

We set each thread process P data, P is the pixel value which required $n = 4B$, if image is in RGBA format. T_i represents any thread in a block, where i is the thread

index. $THREAD_N$ is the total number of threads in each block, $BLOCK_N$ is the block number of each grid, N is the total size of the input data, n 16KB is the size of shared memory of the NVIDIA G80 series cards, so the execution configuration can be set below:

- a) Ti processes P data; $(THREAD_N * P) B < 16KB$;
- b) $BLOCK_N = N / (n * P)$.

It is desirable not to occupy the whole shared memory; some place should be remained to put some special variables. We describe various design strategies for various operations in pyramidal blending algorithm below.

4.1 Construction of Gaussian Pyramid

A sequence of low-pass filtered images G_0, G_1, \dots, G_N can be obtained by repeatedly convolving a small weighting function with an image. The convolution operation is following SIMD paradigm. We apply following two functions in NVIDIA's CUDA. We define proposed strategy for implementation.

(1) *CUDA Gaussian Blur*

The first step is applying 5×5 Gaussian blur filters. We take Gaussian constant equal to 1 in all cases of implementation, the kernel configuration is of 16×16 threads of each block and 32 of blocks on 512×512 pixel image. This kernel configuration is applied to each grid and there are total 32 grids of image size. The convolution is parallelized across the available computational threads where each thread computes the convolution result of its assigned pixels sequentially. Pixels are distributed evenly across the threads. All threads read data from share memory but due to limitation in shared memory data should be moved from global memory to shared memory. Synchronization of the threads can be done by CUDA Synchronized function Blocks. Which will do thread synchronization per block automatically to maintain results.

(2) *CUDA Reduce Operation*

In this operation a sequence of low-pass filtered images G_0, G_1, \dots, G_N can be obtained by repeatedly convolving a small weighting function with an image, which can be worked in grids. With this technique, image sample density is also decreased with each iteration so that the bandwidth is reduced in uniform one-octave steps we first need to reduce the image size by half at each level of pyramid. This implementation can be done in texture memory. The texture memory is used to implement the function using OpenGL graphics library. Standard API will call to execute it in CUDA. Intermediate results of each level images will copied from shared memory to Global memory to implement REDUCE operation as defined in the previous section.

4.2 Construction of Laplacian Pyramid

(1) *Expand Operation*

Expand operation can be achieved by reversing the REDUCE process. This implementation can be done in texture memory. The texture memory is used to implement the function using OpenGL graphics library. Standard API will call to

execute it in CUDA. Intermediate results of each level images will copied from shared memory to Global memory to implement EXPAND operation as defined in the previous section.

(2) *Laplacian of Gaussian*

In order to obtain the band-pass images required for the pyramidal blend we subtract each level of the pyramid from the next lowest level. Because these arrays differ in sample density, it is necessary to interpolate new samples between those of a given array before it is subtracted from the next lowest array. Interpolation can be achieved by reversing the REDUCE process called EXPAND defined above. To implement Laplacian of Gaussian we follow SIMD paradigm. We will use the same thread configuration as we described before. Each thread need the result of Expand operation as described above for each pyramid level so we can get it from Global memory. Intermediate results can be copied from shred memory to Global Memory.



(a)



(b)



(c)

Fig. 3. Pyramidal Blending (a) left image (b) right image (c) Blended panorama

5 Results

In result we have shown pyramidal blending of two images With resolution of 1147×608 . figure 3a, 3b shows left image and right image respectively, figure 3c sows final blended panorama and figure 3d shows time comparison between CPU and GPU implementation.

Table 1. Time comparison

Pyramidal Belding	CPU time(s)	GPU time(s)	Speed up
Combine operation	7.18(s)	2.30(s)	3.13

6 Conclusion

For parallel computing by CUDA, we should pay attention to two points. Allocating data for each thread is important. So if better allocation algorithms of the input data are found, the efficiency of the image algorithms would be improved. In addition, the memory bandwidth of host device is the bottleneck of the whole speed, so the quick read of input data is also very important and we should attach importance to it. Obviously, CUDA provides us with a novel massively data-parallel general computing method, and is cheaper in hardware implementation.

References

1. NVIDIA, CUDA Programming Guide Version 2.3. NVIDIA Corporation: Santa Clara, California Intel, Quad-Core Intel® Xeon® Processor 5400 Series 2008, Intel Corporation: Santa Clara, California (2008), <http://gpgpu.org>, General-Purpose Computation on Graphics Hardware
2. Allard, J., Raffin, B.: A shader-based parallel rendering framework. In: Visualization, 2005, VIS 2005, pp. 127–134. IEEE, Los Alamitos (2005)
3. Neve, D., et al.: GPU-assisted decoding of video samples represented in the YCoCg-R color space. In: Proc. ACM Int. Conf. on Multimedia (2005)
4. Sinha, S.N., Frahm, J.-M., Pollefeys, M., Genc, Y.: Feature tracking and matching in video using programmable graphics hardware. In: Machine Vision and Applications, MVA (2007)
5. Burt, P.J., Andelson, E.H.: A multiresolution blend with application to image mosaics. ACM Transactions on Graphics 2(4) (1983)
6. Adelson, E.H., Anderson, C.H., Bergen, J.R.: Pyramid methods in image processing (1984)
7. Park, S.I., Ponce, S.P., Huang, J., Cao, Y., Quek, F.: Low-Cost, High-Speed Computer Vision Using NVIDIA's CUDA Architecture
8. Yang, Z., Zhu, Y., Pu, Y.: Parallel Image Processing Based on CUDA 978-0-7695-3336-0/08 © 2008. IEEE, Los Alamitos (2008)

Toggle Coverage for ALU Using VHDL

D. Venkat Reddy¹, Ch.D.V. Paradesi Rao², and E.G. Rajan³

¹ MGIT, Gandipet, Hyderabad-500028
dasar_i_reddy@yahoo.com

² Arora Engg. College, Bhongir, AP

³ Pentagram Research Center, Jubilee Hills, Hyderabad, AP

Abstract. Designing modern circuits comprised of millions of gates is a very challenging task. Therefore new directions are investigated for efficient modeling and verification of such systems. In this paper, we first present the concept of modeling multiple valued ALU. We demonstrate that the approach allows for efficient simulation of complex multiple valued logic systems. Secondly, we show how VHDL can be used to ensure functional correctness. A generalization of binary toggle coverage for the multiple valued logic domains is presented and evaluated. As a test case, a scalable multiple valued logic arithmetic unit is modeled and experimental results for multiple valued logic toggle coverage are given.

Keywords: Ternary Logic Gates, Multiple Valued Logic, FPGA.

1 Introduction

The rapid development of complex systems requires a reliable modeling and verification approach. Currently, a gap exists in terms of abstraction between the languages used to describe hardware at the Register Transfer Level (RTL) and design languages used to describe the first reference design. Verilog and VHDL are typically used to describe hardware at the RTL and a different language (C, C++) is commonly employed for the first reference design. This results in a significant amount of effort since the same code has to be written at two or more different levels of abstraction in multiple programming languages. A new language, SystemVerilog [2], aims to bridge this gap by extending Verilog to higher levels of abstraction for architectural and algorithmic design and for advanced verification [4, 6]. There has been a significant past effort in the development of methods for designing Multiple Valued Logic (MVL) circuits. Modeling circuits based on MVL instead of the binary domain can help during verification [10]. Thus, a development environment that supports modeling of MVL circuits is required. A first approach in this direction was introduced in [5] where a package to model ternary circuits in VHDL was provided. In [1] a modeling platform using SystemC was presented. In addition to modeling MVL circuits, their correct functional behavior must also be ensured. Therefore in [1] pure simulation is used but the completeness of the verification process is not discussed.

For simulation-based design validation, an estimate of the functionality that is exercised can be given using coverage metrics [8]. A metric based on the circuit

structure and one that is well known for the binary case is toggle coverage [7]. The basic idea is to check whether the value of an internal net toggles between all possible values. For the MVL domain this metric has not been studied previously.

In this paper we present a concept for modeling MVL circuits in VHDL. The concept is demonstrated for a scalable MVL Arithmetic Logic Unit (ALU). To determine the quality of the verification process we introduce a generalization of toggle coverage for the MVL domain. We show that two types of toggle coverage should be distinguished if the underlying logic is MVL.

2 MVL Modeling in VHDL

2.1 Multiple Valued Logic Circuit

MVL is a non-binary logic that involves switching between more than 2 values. We will assume that MVL primitive devices will be limited to 2-input/single-output functions. For example, a logic function with radix-3 is one that has two inputs that can assume three values (i.e. 0, 1, or 2) and generates one output signal that can have one of these three values. Symbolically, a logic function based upon a single radix-3 device can be graphically depicted as shown in Figure 1. Table 1 shows an example truth table of such a device.

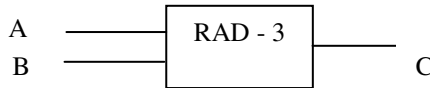


Fig. 1. Logic function with radix 3

Table 1. Truth table of a function with radix 3

A/B	0	1	2
0	1	0	2
1	0	2	1
2	2	1	0

The example function in Table 1 is commutative, and hence it would not matter if we exchange the 'A' and 'B' inputs. MVL circuits can be modeled as graphs where the edges correspond to nets and the vertices correspond to storage elements or MVL basic gates [3]. All of the basic MVL gates can be used as an operator in higher level descriptions of a circuit. More complex constructs can be defined based on the MVL gates and can be used in the high level description as well.

Given r as the radix, the operators (which correspond to basic gates) are defined on a set of $O = \{0, \dots, r-1\}$. We chose a functional complete set of operators [1], all of them having one output and one or two operands x and y ($x, y \in O$) as their inputs. The operators and their functionalities are shown in Table 2.

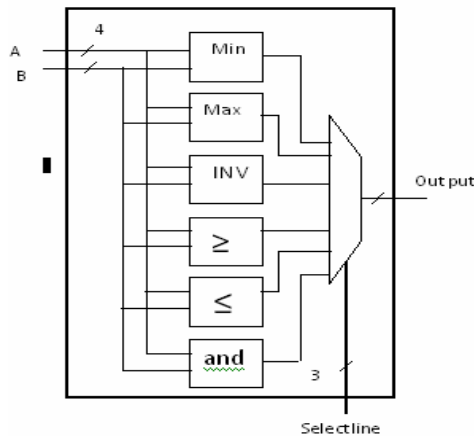
Table 2. Table for function and Description

Operators	Function
MIN(x,y)	Minimum {x,y}
MAX(x,y)	Maximum {x,y}
INV(x,y)	$(r-1)-x$
Equal(x,y)	Equal {x,y}
Grater than(x,y)	1 if $x \leq y$ 0 otherwise
Smaller than(x,y)	1 if $x \leq y$ 0 otherwise
Logical and(x,y)	$C = x \text{ and } y$

In the binary case MIN, MAX and INV correspond to AND, OR and INV, respectively

2.2 Case Study – MVL ALU

A scalable MVL ALU circuit for different radices is implemented in a VHDL description. The block diagram of the MVL ALU circuit is shown in Fig. 2

**Fig. 2.** ALU circuit

The ALU has four inputs: clock, two MVL operand inputs A and B, and the control input select. The ALU can perform the INV, MIN, MAX, addition and multiplication operations. The output is stored in the register named Res. The width of the operands A and B is equal and the result bus Res has double width to accommodate the result of a multiply operation. The VHDL implementation of the ALU is scalable since the width and radix are separately defined and can be easily varied.

3 Toggle Coverage

To understand the theory of Toggle coverage let us assum a MVL vector called “A”, we have checked if each element has taken all the legal values according to the set {0, 1, 2, 3, 4}. For example, if the first element of A, denoted as A[0], has taken the values 0, 1 and 3 during simulation, then the value of type I toggle coverage for A[0] is 3/5 or 60%. For the complete vector the coverage is determined as an average value of coverage computed over all elements of A.

For type II toggle coverage, consider the following example. Again we have an MVL vector called “A” with radix 5. For the first element A[0] of A one would check that there were transitions between all values of the set {0, 1, 2, 3, 4}; i.e. 0→1, 0→2, 0→3, 0→4, 1→0,..., 4→3. In general radix*(radix-1) transitions have to be considered. The transitions which need to be checked for the example are shown in Table 3, where ‘T’ indicates a transition of interest whereas ‘X’ indicates a transition that does not give useful information. Hence, if A[0] has made 10 out of the 20 possible transitions, we say that the type II toggle coverage is 50% for A[0]. The coverage for the whole vector A is once again computed as an average over the coverage value of each element of A.

Table 3. Toggle coverage type I

Loc	0	1	2	0	1	2	0	1	2	0	1	2
A[0]	1	1	1	1	...
A[1]	...	1	...	1	1	1
A[2]	1	1	1	1
A[3]	1	1	1	1	...
COV	1	1	1	1	1	1	1	1	1	1	1	...

To demonstrate our implementation of toggle coverage type I using an example, we consider an MVL vector A of width equal to 4 and radix equal to 3. We check if each element has taken all the legal values according to the set {0, 1, 2}. To perform this computation for vector A, we create an array of size width of A * radix. For example the size of the array is 4*3 = 12. Say, we name this array “cov”. If the element of A takes a value, then we set a 1 at location of cov. Hence, if the vector A takes the following four values successively during simulation: [0, 1, 2, 0], [1, 0, 2, 1], [1, 2, 0, 2], [1, 0, 0, 1], then the toggle coverage of type I for vector A is 11/12, or 91% . It should be noted here that while computing the coverage of the result vector the size of the array to be created will be twice (i.e. width of A * radix * 2) since the result vector is twice as big as A.

We now discuss our implementation of toggle coverage type II again with the example of an MVL vector A and radix 3. In this case 3*2=6 transitions have to be considered between all values of the set {0, 1, 2}; i.e. 0→1, 0→2, 1→0, 1→2, 2→0, 2→1, for each element of A. Hence, for the vector A of width 4, 6*4=24 different

Table 4. Toggle coverage Type II

	0	0	1	1	2	2	0	0	1	1	2	2	0	0	1	1	2	2	
	1	2	0	2	0	1	1	2	0	2	0	1	1	2	0	2	0	1	
A[0]	1	1	1	
A[1]	1	1	1
A[2]	1	1	1
A[3]	1	1	1
COV	1	1	1	1	1	1	1	1	1	1

transitions need to be checked in order to compute the toggle coverage of type II. If the vector A takes the same four values as before: [0, 2, 1, 0], [2, 0, 1, 1], [1, 1, 0, 0], [0, 2, 1, 2] then the according array content of toggle coverage type II for A is illustrated in Table 4.

After each clock cycle we compare the new value of A with the previous value of A. If the value has changed we set a 1 to the appropriate position, and keep the value of the previous cycle as well. This calculation is more exhaustive in the sense that transitions and not just instantaneous values are considered. We observe that the coverage for vector A is $10/24=42\%$. As expected this value is lower than the toggle coverage value for type I (91%), since type II of coverage is more exhaustive.

Again the presented formulation of the function is general because the width and radix of an MVL vector for which the coverage is computed can be arbitrary.

4 Simulation Result

In this section MVL coverage is computed for different instances of the MVL ALU.

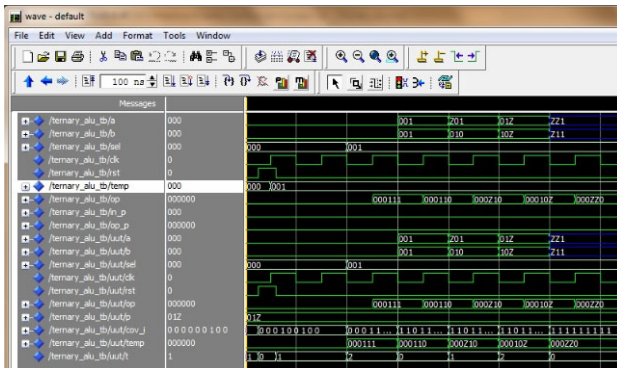


Fig. 3. Simulation results

4.1 Coverage Estimation

In the following experiments the radix of the ALU is 3 and the width of the input vectors A and B is 4. During the simulation, random values for all inputs of the ALU

are generated using VHDL constrained based randomization technique. Thus, the value of each input is chosen with a uniform distribution out of the legal value set. The results of the operations MIN, MAX, INV, ADD, and MULT are referring to the ALU circuit in Fig. 2 as nets 1-5, respectively. The width of all these nets is 8 (width*2) since the largest result is the product value of the multiplication unit that determines the size of the other inputs for the multiplexer. In the following we analyze the coverage for these nets. All coverage numbers have been obtained by averaging the results over 100 runs. In Fig. 4, the results for toggle coverage type I are shown. As expected, the coverage for the vectors for operations MIN, MAX and INV did not exceed 66%, since only the 4 lower digits of the result vector are affected and the remaining digits do not toggle. In case of the ADD operation 71% were achieved since 5 lower digits of the result vector toggle. However, in case of the MULT operation the coverage is 100% since all digits of the vector toggle. Similarly, in Fig. 5 we show the toggle coverage of type II.

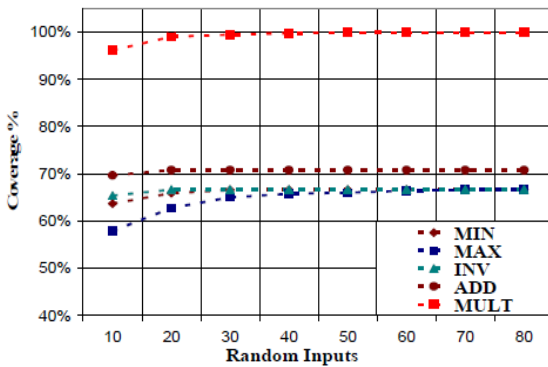


Fig. 4. Toggle coverage type I vs. number of inputs, vector width = 4

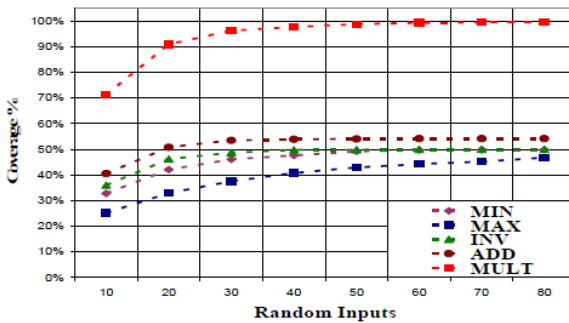


Fig. 5. Toggle coverage type II vs. number of inputs, vector width = 4

Both graphs show that the coverage converges very fast. As expected for the toggle coverage type II more random inputs are necessary to achieve convergence.

5 Conclusion and Future Work

In this paper we first presented a concept for modeling MVL circuits using VHDL that allows for easy scaling of operand size (in terms of digits) and radix value. A scalable MVL ALU has been implemented and simulated efficiently. To determine the quality of the verification process we generalized the concept of toggle coverage for MVL. It was shown that two types of toggle coverage must be distinguished for the MVL domain. In the experimental results, the convergence of both coverage types for the MVL ALU is empirically shown.

Future directions in this research include developing a general flow to automate the code instrumentation of an MVL VHDL design.

References

- [1] Große, D., Fey, G., Drechsler, R.: Modeling Multi-Valued Circuits in SystemC. In: International Symposium on Multi Valued Logic, pp. 281–286 (2003)
- [2] The SystemVerilog Homepage, <http://www.systemverilog.org>
- [3] Muzio, J., Wesselkamper, T.: Multiple-Valued Switching Theory. Adam Hilger, Bristol (1986)
- [4] Accellera: SystemVerilog 3.1 Accellera's extensions to Verilog, Npa, CA, pp. 1–2 (2003)
- [5] Rozon, C.: On the Use of VHDL as a Multi-Valued Logic Simulator, Royal Military College, Kingston, Ontario, Canada. IEEE, Los Alamitos (1996)
- [6] Cohen, B., Venkataramanan, S., Kumari, A.: SystemVerilog Assertions Handbook. VhdlCohen Publishing, Los Angeles (2005)
- [7] Drako, D., Cohen, P.: HDL Verification Coverage. EETimes, Global news for creators of technology (1998)
- [8] Chokler, H., Kupferman, O., Vardi, M.Y.: Coverage Metrics for Formal Verification. In: Geist, D., Tronci, E. (eds.) CHARME 2003. LNCS, vol. 2860, pp. 111–125. Springer, Heidelberg (2003)
- [9] Accelera SystemVerilog 3.1a reference manual, Extension to Verilog, <http://www.accellera.org/home>
- [10] Drechsler, R.: Using World-Level Information in Formal Hardware Verification. Automation and Remote Control 65(6), 963–977 (2004)
- [11] Iguchi, Y., Sasao, T., Matsuura, M.: On Designs of Radix Converters Using Arithmetic Decompositions - Binary to Decimal Converters. In: 37th International Symposium on Multiple-Valued Logic, ISMVL 2007, p. 32 (2007)
- [12] Ito, T., Kameyama, M.: Universal VLSI Based on a Redundant Multiple-Valued Sequential Logic Operation. In: 37th International Symposium on Multiple-Valued Logic, ISMVL 2007, pp. 39–44 (2007)

Design and Implementation of Voltage Control Oscillator (VCO) Using 180nm Technology

M.R. Halesh¹, K.R. Rasane², and H. Rohini³

¹ Electronics & Communication, B.V. Bhoomaraddi College of Engg., & Tech,
Hubli, Karnataka, India

² Electronics & Communication, K.L.E.C.E.T, Belgaum, Karnataka, India

³ Electronics & Communication, B.V. Bhoomaraddi College of Engg., & Tech.,
Hubli, Karnataka, India

{haleshmr, rohini_sh}@bvb.edu,
kruparasane@hotmail.com

Abstract. Voltage Control Oscillator (VCO) is an integral part of many electronic applications like PLL, clock generation in microprocessors & carrier synthesis in cellular telephones etc. Such applications require different topologies which gives robust high performance. Consequently, VCO design in CMOS technology continues to pose interesting challenges. This paper presents the design of Voltage Control Ring Oscillator with the oscillation frequency up to 1 GHz. The circuit is implemented using Cadence tool in 0.18 μ m CMOS technology (UMC180) with 1.8V supply. The designed VCO is generating a frequency of 1.06GHz over a temperature range from -40°C to 125°C, & the linearity is achieved over a range of frequency from 970MHz to 1.03GHz.

Keywords: VCO, PLL, DRC, LVS.

1 Introduction

The Phase Lock Loop (PLL) is a critical component in many high speed systems as it provides the timing basis for the functions such as clock control, data recovery and synchronization. The VCO is perhaps the integral element of the PLL. A CMOS VCO can be built using ring structure, relaxation circuit or an LC resonant circuit. The LC design has the best phase noise and frequency performance, but their tuning range is relatively small. However adding high quality inductors to a CMOS process flow increases the cost and complexity of the chip.

Ring oscillator, on the other hand can be built in any standard CMOS process and may require less die area than LC design. This paper presents a design of ring oscillator with a conventional architecture, which requires connecting an odd number of inverters and feedback from the output of last stage to the input of the first stage. The oscillation frequency is determined by the number of stages and the delay in each stage which is very small for inverters. If the delay is voltage controllable, then a VCO with variable-frequency output is obtained. The proposed ring oscillator is used for the application of PLL as a communication circuit.

2 Design Implementation

A conventional VCO as shown in figure 1, is realized by N stages of inverters (N is an odd number), with a control mechanism of the current passing through these inverters. Usually inverter current mirror circuit is used to generate a current in each delay stage.

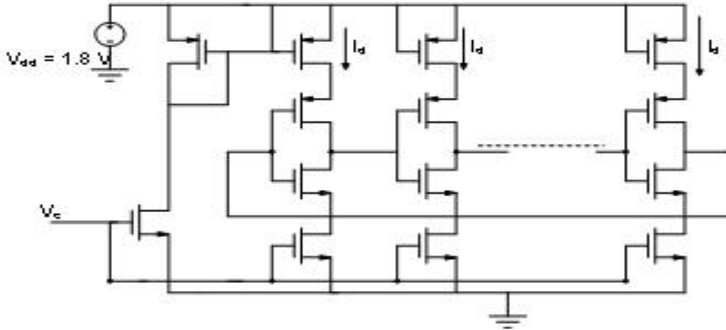


Fig. 1. Conventional VCO

The frequency of oscillation can be found as

$$f_{osc} = \frac{1}{2N\tau} \tag{1}$$

Where τ is the delay of each stage.

The gain of the amplifier (inverter) from small signal model in figure 2 is given as

$$A = g_m r_{ds} \tag{2}$$

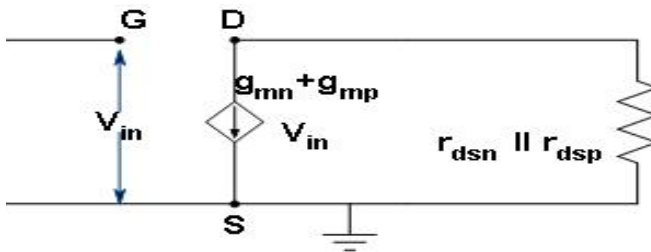


Fig. 2. SSM of an Inverter

Yalcin Alper Eken, and John P. Uyemura [1], compared the Differential architecture with the LC resonator circuit (relaxation circuit) and the LC circuits are usually implemented for a high frequency. In this paper an attempt has made to achieve high frequency using CMOS circuit. Some of the characteristics like area and power is improved in CMOS circuit compared to LC oscillator.

3 Feedback and Stability

Figure 3 represents the block schematic of a feedback amplifier system, with transfer gain A and feedback network β [5], and a mixing circuit connected to form a closed loop. The amplifier provides an output signal V_o as consequence of external signal V_{in} applied directly to amplifier input terminal.

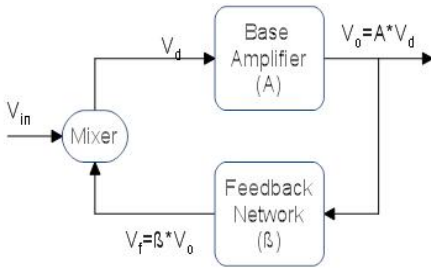


Fig. 3. Feedback amplifier system

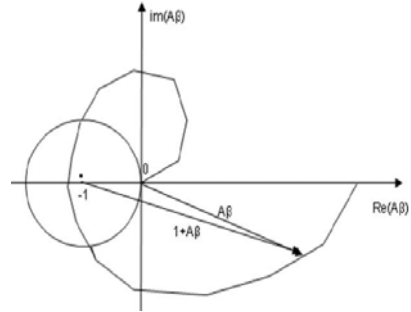


Fig. 4. The locus of $1 + A\beta = 1$ is a circle Of unit radius, with center $-1 + j0$

The overall gain of the system is found to be

$$A_f = A / (1 + A\beta) \tag{3}$$

If $1 + A\beta = 0$ (i.e., $-A\beta = 1 \angle 0^\circ$) this is called Barkhausen’s criterion for sustained oscillation. An improved version of the Barkhausen criterion was given by Nyquist for understanding the stability of amplifier using Loop gain[5]. If $-A\beta > 1 \angle 0^\circ$, the amplifier becomes unstable, and the output sees a growing oscillation. The criteria for positive or negative feedback may also be represented in the complex plane. From Figure 4, we see that $1 + A\beta = 1$, represents a circle of unit radius, centered at $(-1, 0)$. Thus that portion of $A\beta$ curve that falls outside the unit circle indicates negative feedback, and the polar plot within the unit circle indicates positive feedback. For oscillation, the polar plot should cross the critical point $(-1, 0)$, and the frequency of oscillation is the point where the curve crosses the negative real axis.

4 The Current Starved VCO

The current starved VCO is shown in Figure 5. Its operation is similar to the ring oscillator [7]. MOSFETs M_2 and M_3 operate as an inverter, while MOSFETs M_1 and M_5 operate as current sources. The MOSFETs M_5 and M_6 drain currents are the same and are set by the input control voltage. The currents in M_5 and M_6 are mirrored in each inverter/current source stage. Consider the simplified schematic of one stage of the VCO shown in Figure6, to derive the design equations. The total capacitance on the drains of M_2 and M_3 is given by

$$\begin{aligned} C_{tot} &= C_{out} + C_{in} \tag{4} \\ &= C_{ox} ' (W_p L_p + W_n L_n) + 1.5 C_{ox} ' (W_p L_p + W_n L_n) \\ &= 2.5 C_{ox} ' (W_p L_p + W_n L_n). \end{aligned}$$

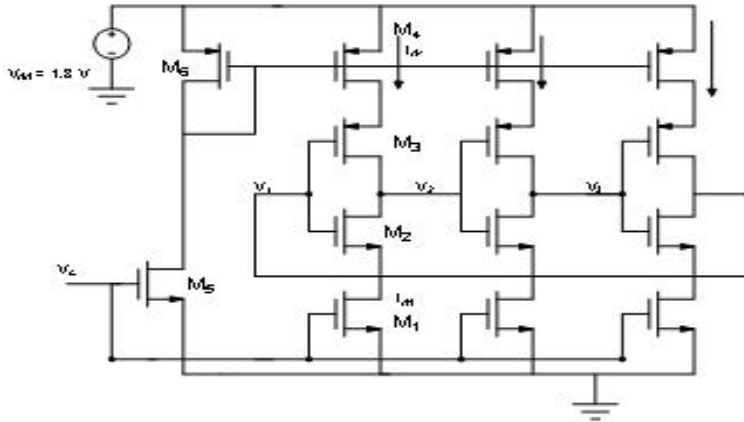


Fig. 5. Current starved VCO

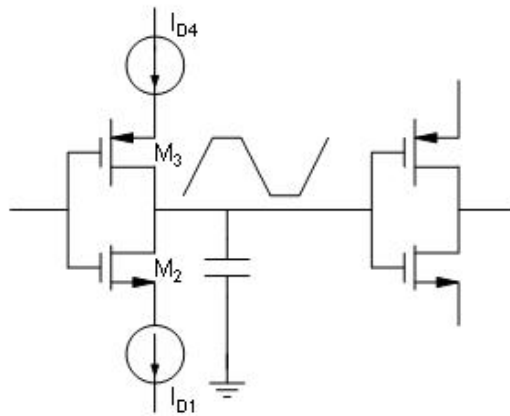


Fig. 6. Simplified view of single stage of current-starved VCO

The time takes to charge C_{tot} from zero to V_{SP} with constant current I_{D5} is given by

$$t_1 = C_{tot} (V_{SP}/I_{D4}) \tag{5}$$

The time taken to discharge from V_{DD} to V_{SP} is given by

$$t_2 = C_{tot} (V_{DD}-V_{SP})/I_{D1} \tag{6}$$

If we set $I_{D5}=I_{D1}=I_D$ (which is $I_{D_{centre}}$ when $V_{inVCO} = V_{DD}/2$) then the sum of t_1 and t_2 is simply

$$t_1 + t_2 = (c_{tot} * V_{DD}) / I_D \tag{7}$$

The oscillation frequency of the current starved VCO for N (an odd number) of stages is

$$f_{osc} = 1 / N(t_1 - t_2)$$

$$f_{osc} = I_D / (N * C_{tot} * V_{DD}) \quad (8)$$

This is equal to

$$f_{osc} = f_{centre} \text{ at } V_{inVCO} = V_{DD}/2 \text{ and } I_D = I_{center} \quad (9)$$

5 Physical Design

The process of translating the net-list into Silicon wafer requires an accurate map of the circuit with details of device size, their layout and interconnections. The final layout of VCO is shown in Fig 6, consuming $697 \text{ } (\mu\text{m})^2$ of silicon areas.

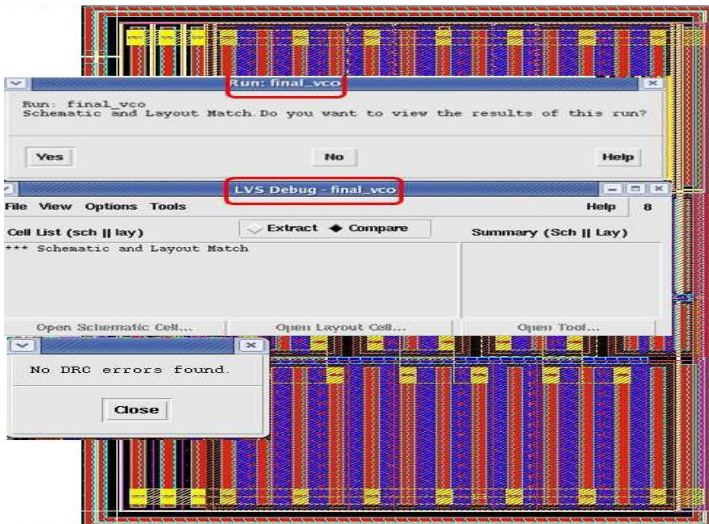


Fig. 6. Implementation of VCO with its validation result

6 Results and Discussions

Transient Analysis

The transient analysis of the circuit is depicted in figure 7a showing generated frequency of 1.06 GHz in a typical corner. The delay between each stage is measured to be 326ps, 306ps, and 304ps between V_1 to V_2 , V_2 to V_3 and V_3 back to V_1 respectively, from the circuit shown in the Figure 5. So that the total delay is obtained to be 936ps and so we get a frequency of 1.068GHz. This circuit was simulated for all the process corners with 10% variation in power supply and the temperature variation from -40°C to 125°C .

Table 1. Comparison with different Architectures

	Architecture-1(Ref-1)	Architecture-2 (Ref-9)	Architecture-3 (Ref-2)	Our paper
Technology	180-nm CMOS Technology	0.35 μ m - 2P3M-COS Technology	0.8 μ m High Voltage CMOS/DMOS Tech with 5V	UMC180nm Technology with 1.8V
Frequency Range	3 – Stages 5.16GHz - 5.93 GHz	3 – Stages 1.25 GHz	3 – Stages 13 Hz - 407 MHz	3 – Stages 970 MHz - 1.03 GHz
Phase Noise Measured	-99.5 dBc/Hz at 1-MHz offset from a 5.79-GHz center frequency.	-20dBc/Hz at 100kHz.	---	----

Table 2. The pre-layout & post-layout simulation results

Corners	Supply voltage in Volts	Frequency range in GHz (prelayout)	Frequency range in GHz (Postlayout)
Typical	1.8	1.063	1.047
	1.68	0.971	0.958
	1.92	1.151	1.132
Slow	1.8	1.024	0.935
	1.68	0.857	0.851
	1.92	0.945	1.015
Fast	1.8	1.180	1.154
	1.68	1.083	1.061
	1.92	1.271	1.241
Slow nMOS & Fast pMOS (Snfp)	1.8	1.144	1.040
	1.68	0.968	0.953
	1.92	1.058	1.123
Fast nMOS & Slow pMOS (Fnsp)	1.8	1.063	1.050
	1.68	0.970	0.959
	1.92	1.152	1.136

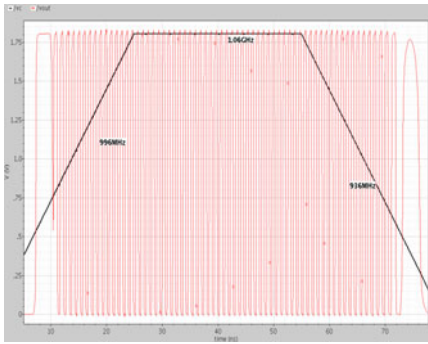


Fig. 7a. Transient Analysis - variation of frequency with control voltage

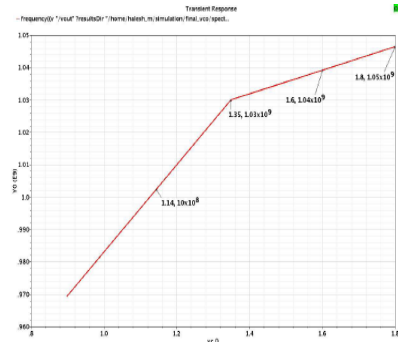


Fig. 7b. Linearity variation of frequency with control voltage

7 Conclusion

A voltage controlled oscillator is designed and implemented for the frequency of 1GHz. The parasitic are extracted and the back annotated circuit is simulated for all the process corners, for a temperature ranging from -40°C to 125°C . The transient analysis shows the obtained frequency is 1.06GHz in a typical corner and 860MHz in a worst corner with 10% supply variation. The result in figure 7b shows linear variation of output frequency with respect to control voltage over a frequency range of 970MHz to 1.03GHz. The implemented architecture works for the frequency of 1.06GHz. Further high frequency can be achieved by varying the length of the active device & current flowing in the delay stage.

References

1. Eken, Y.A., Student Member, IEEE, Uyemura, J.P.: A 5.9-GHz Voltage-Controlled Ring Oscillator in 0.18- μm CMOS. *IEEE Journal of Solid-State Circuits* 39(1) (January 2004)
2. Chebli, R., Zhao, X., Sawan, M.: A Wide Tuning Range Voltage-Controlled Ring Oscillator dedicated to Ultrasound Transmitter
3. Sicard, E., Bendhia, S.D.: Basics of CMOS Cell Design
4. Hajimiri, A., Limotyrakis, S., Lee, T.: Jitter and Phase Noise in Ring Oscillators. *IEEE Journal of Solid-State Circuits* 34(6) (June 1999)
5. Millman, J., Halkias, C.C.: *Electronic Devices and Circuits*
6. Kang, S.-M., Leblebici, Y.: *CMOS Digital Integrated Circuits- Analysis and Design*
7. Baker, R.J., Li, H.W., Boyce, D.E.: *CMOS Circuit Design, Layout and Simulation*. In: Tewksbury, S.K. (ed.). *IEEE Press Series on Microelectronic System*, Series edn.
8. Razavi, B.: *Design of Analog CMOS Integrated Circuits*
9. Xiao, L., Liu, W., Yang, L.: State Key Lab of ASIC and System, Fudan University, Shanghai 201203, P. R. China
10. Allen, P.E., Holberg, D.R.: *CMOS Analog Circuit Design*, 2nd edn.

Design of Robust PID Controller for Flexible Transmission System Using Quantitative Feedback Theory (QFT)

Mukesh D. Patil* and Kausar R. Kothawale

Department of Electronics Engineering,
Ramrao Adik Institute of Technology, Nerul,
Navi Mumbai, 400 706, India

Tel.: +91-22-27709574

mdpatil@sc.iitb.ac.in,

kausarkr@yahoo.com

<http://www.rait.ac.in>

Abstract. Many practical systems are characterized by high uncertainty which makes it difficult to maintain good stability margins and performance properties for closed loop system. In case of conventional control, if plant parameter changes we can not assure about the system performance hence it is necessary to design robust control for uncertain plant. Among the various strategies proposed to tackle this problem, Quantitative Feedback Theory (QFT) has proved its superiority especially in the face of significant parametric uncertainty. The feature of QFT is that it can take care of large parametric uncertainty along with phase information. For the purpose of QFT, the feedback system is normally described by the two-degrees-of freedom structure. A PID controller is a generic control loop feedback mechanism widely used in industrial control systems. A flexible transmission system that has uncertainties in the frictional losses, stiffness of spring and inertial constant is presented using robust PID controller.

Keywords: Robust Control, QFT, PID controller, Flexible transmission system.

1 Introduction

In recent years, control systems have assumed an increasingly important role in the development and advancement of modern civilization and technology. Control systems are an integral component of any industrial society and are necessary for the production of goods. We find control systems in all sectors of industry, such as quality control of manufactured products, automatic assembly line, machine tool control, space technology, power systems [1]. Various control strategies are used for design of control system depending on plant model. If

* Corresponding author.

the design performs well for substantial variations in the dynamics of the plant from the design values, then the design is robust. Robust control deals explicitly with uncertainty in its approach to controller design. QFT is robust control method which deals with the effects of uncertainty systematically. QFT is a graphical loop shaping procedure used for the control design of either Single Input Single Output (SISO) or Multiple Input Multiple Output (MIMO) uncertain systems including the nonlinear and time varying cases [2] [3]. QFT, developed by Isaac Horowitz is a frequency domain technique utilizing the Nichols chart (NC) in order to achieve a desired robust design over a specified region of plant uncertainty. In comparison to other robust control methods, QFT offers a number of advantages. For the purpose of QFT, the feedback system is normally described by the two-degrees-of freedom structure. A proportional-integral-derivative controller (PID controller) is a generic control loop feedback mechanism widely used in industrial control systems. A PID controller attempts to correct the error between a measured process variable and a desired set-point by calculating and then outputting a corrective action that can adjust the process accordingly.

2 Some Preliminaries

2.1 Basic of PID Controller

Proportional Integral Derivative control is the most common type of controller used in the industrial control system. A PID controller in parallel form (also known as standard form or non-interacting form), as shown in figure 1 has the control equation as given below [4]:

$$u = K_p(1 + \frac{1}{T_i s} + T_d s)e \tag{1}$$

Where, e is the error and u is the control effort. The equation 1 can be rewritten as:

$$u = K_p(\frac{T_i T_d s^2 + T_i s + 1}{T_i s})e \tag{2}$$

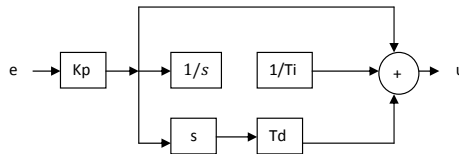


Fig. 1. The Parallel form PID Controller

2.2 Quantitative Feedback Theory

Consider a two degree freedom feedback system configuration (see Fig 2), where $P(s)$, $G(s)$ and $F(s)$ are uncertain linear time-invariant plant, the controller and pre-filter to be designed respectively.

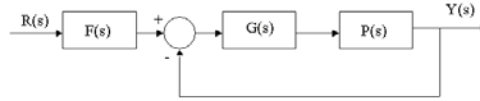


Fig. 2. The Two Degree-of-Freedom Structure in QFT

The open loop transmission function is defined as

$$L(s) = G(s)P(s) \tag{3}$$

and the nominal open loop transmission function is

$$L_0(s) = G(s)P_0(s) \tag{4}$$

The objective in QFT is to synthesize $G(s)$ and $F(s)$ such that the various stability and performance specifications are met for all $P(s) \in \mathcal{P}$. In general following specifications are considered in QFT. [2]:

- Robust stability margin

$$\left| \frac{L(j\omega)}{1 + L(j\omega)} \right| \leq W_s$$

- Robust tracking performance

$$|T_L(j\omega)| \leq \left| \frac{F(j\omega)L(j\omega)}{1 + L(j\omega)} \right| \leq |T_U(j\omega)|$$

- Robust input disturbance rejection performance

$$\left| \frac{G(j\omega)}{1 + L(j\omega)} \right| \leq W_{d_i}(w)$$

- Robust output disturbance rejection performance

$$\left| \frac{1}{1 + L(j\omega)} \right| \leq W_{d_o}(w)$$

In practice, the objective is to satisfy the given specifications over a finite design frequency set Ω . The design procedure which is to be followed for applying QFT robust design technique is as follows [3]:

- Synthesize the desired tracking model
- Specify the plant models that define the region of plant parameter uncertainty

- Obtain the plant templates at specified frequencies that pictorially describe the region of plant parameter uncertainty on the Nichols Chart.
- Select the nominal plant transfer function $P_0(s)$.
- Determine the stability contour on the Nichols Chart.
- Determine tracking and optimal bounds on the Nichols Chart.
- Synthesize the nominal loop transmission function $L_0(s) = G(s)P_0(s)$ that satisfies all the bounds and stability contour.
- Synthesize the pre-filter $F(s)$.

3 Control System for Flexible Transmission System

The flexible transmission system consists of three horizontal pulleys connected by two elastic belts. The schematic diagram of the system is shown in fig 3. The first pulley is driven by a dc motor whose position is controlled by local feedback. Since the dynamic of this feedback loop is much faster than that of the mechanical parts, it can be neglected in the analysis of the system. The objective is to control the position of the third pulley which may be loaded with small disks. The system input is the reference for the axis position of the first pulley. The system is characterized by two low damped vibration modes (with damping factors of less than 0.05), subject to large variations in the presence of load 5.

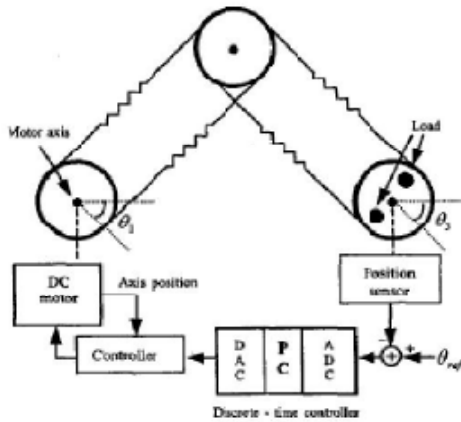


Fig. 3. Schematic Diagram of Flexible Transmission System

Fig. 4 gives the amplitude of the time characteristic of the identified continuous time models for three different loadings, no load, half load and full load. A variation about 100percent of the frequency of the first vibration mode occurs when passing from full loaded case to unloaded case. The system is very oscillatory so the challenge is to obtain robust controller that gives desired overshoot.

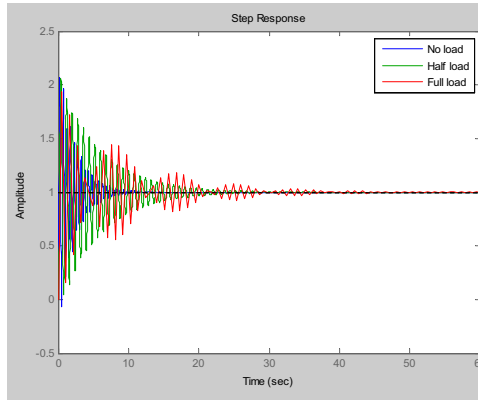


Fig. 4. Time characteristic diagram of flexible transmission system for various loads

The continuous plant is described by the following transfer function

$$G(s) = \frac{\Theta_3(s)}{\Theta_1(s)}$$

$$G(s) = \frac{(2kr^2)^2}{(J_3s^2 + fs + 2kr^2)(J_3s^2 + fs + 4kr^2) - 2kr^2}$$

The transfer functions of the models for three load cases are given below.

No Load

$$G(s) = \frac{11.56}{714 * 10^{-7}s^4 + 1225 * 10^{-7}s^3 + 0.08607s^2 + 0.07395s + 11.56} \quad (5)$$

Half Load

$$G(s) = \frac{11.56}{2159 * 10^{-7}s^4 + 2458 * 10^{-7}s^3 + 0.2017s^2 + 0.07395s + 11.56} \quad (6)$$

Full Load

$$G(s) = \frac{11.56}{3604 * 10^{-7}s^4 + 3690 * 10^{-7}s^3 + 0.3173s^2 + 0.07395s + 11.56} \quad (7)$$

The uncertain plant containing all the three transfer functions is defined as

$$G(s) = \frac{11.56}{P_1s^4 + P_2s^3 + P_3s^2 + 0.07395s + 11.56} \quad (8)$$

where,

$$P_1 \in [714 * 10^{-7}, 3604 * 10^{-7}]; \quad P_2 \in [1225 * 10^{-7}, 3690 * 10^{-7}]; \quad P_3 \in [0.08607, 0.2017]$$

Performance specifications:

- Stability margin specification: Gain margin ≥ 4.5 dB, phase margin $\geq 45^\circ$.
- Tracking specifications: $0.8 \leq \text{rise time} \leq 1.5$ sec, and overshoot ≤ 10 percent.

The nominal loop transmission function $L_0(s) = G(s)P_0(s)$ that satisfies all the bounds and stability contour is synthesized using "lpshape" environment of MATLAB. (see figure 5)

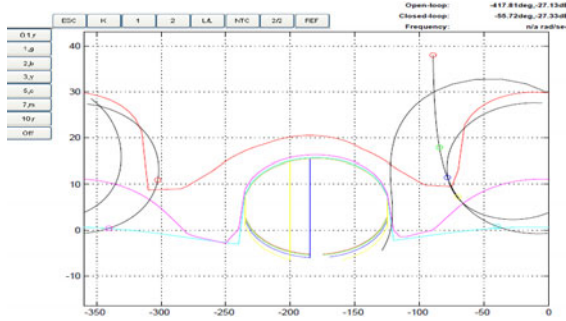


Fig. 5. The Nominal Loop Transmission Function By Synthesizing The Controller

Corresponding controller transfer function obtained for flexible transmission system is as follows:

$$G(s) = 7.978 \frac{0.0246s^2 + 0.0888s + 1}{s} \tag{9}$$

Comparing above transfer function with equation 2, we get PID parameters as

$$K_p = 0.7084; \quad T_i = 0.0888; \quad T_d = 0.277$$

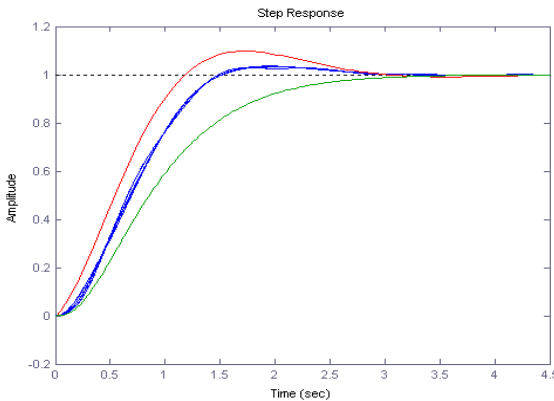


Fig. 6. Closed loop responses of all plant sets for step input satisfying desired tracking specifications

The designed controller and prefilter are simulated using MATLAB environment to verify the performance of designed controller and prefilter. The designed controller and prefilter are verified for set of uncertain plant. The step response of uncertain plant sets is shown in fig. 6. From the result it is observed that all the plant sets satisfies desired time domain specifications. The step response of all plant sets lies in between upper and lower tracking psecifications.

4 Conclusion

The conventional control systems, Lead and lag compensators are used quite extensively in control. For uncertain plant, the robust controller has to be designed. The modern control systems are H_2 , H_∞ and μ -synthesis can't handle large uncertainty and applicable to single input single output (SISO) LTI systems. In the above methods discussed, large parameter uncertainty can't be handled and most of the methods are applicable for designs of SISO LTI systems. The drawbacks of conventional and modern control theory are eliminated by classical control theory based methods which can handle large parameter uncertainty and works in frequency domain called as QFT. Using QFT method we designed a PID controller for an industrial application i.e. flexible transmission system. With design of prefilter we get the tracking requirements.

References

1. Nagrath, L.J., Gopal, M.: Control Systems Engineering. New Age International (P) Limited (2002)
2. Horowitz, I.: Quantitative Feedback design Theory, QFT, vol. 1. QFT Press, Colorado (1993)
3. Houpis, C.H.: Quantitative Feedback Theory: Fundamentals and applications. Taylor and Francis Group Publication, Taylor (2006)
4. Zolotas, A.C., Halikias, G.D.: Optimal Design of PID controllers using the QFT method. In: IEEE Proceedings, Control Theory Appl., vol. 146, pp. 585–589 (1999)
5. Kebriaei, H., Rahimi-Kian, A.: Robust Control of Interval Systems Using a Pole Placement Design. In: IEEE International Conference on Control and Automation, Guangzhou, China, pp. 2939–2944 (2007)

Flatness Based Formation Control of Non Holonomic Vehicle

Ch. Venkatesh*, Sunil K. Surve, and N.M. Singh

Veermata Jijabai Technological Institute(V.J.T.I),
Electrical Engineering Department, Mumbai
surve@frcrce.ac.in, chanti.venky47@gmail.com, nmsingh59@gmail.com

Abstract. In this paper, we use a different viewpoint of flatness which uses a coordinate change based on a Lie-Backlund approach to equivalence in developing flatness-based feedback linearization and its application to the design of controller for formation of Non Holonomic Vehicle. The flat output provides the framework to derive the endogenous feedback compensator, which can result in a constant linear controllable system, for a given nonlinear system. The key contribution of paper is to propose and develop a novel strategy of flatness-based feedback linearization to enhance the stability of formation of the Non Holonomic Vehicle. The proposed flatness-based controller is validated using MATLAB simulation. The simulation results shows that the individual systems maintains the specified geometric pattern and tracks the desired trajectory.

Keywords: Endogenous Feedback, Flatness, Formation, Non Holonomic Vehicle.

1 Introduction

Robot team formation means that the group of robots forms a specific geometric pattern and maintains it while moving. robotic applications, there are several approaches to formation control referred in the literature, namely leader-follower, virtual structures, behavioral-based. In the leader-follower approach, one of the agent is designated as leader, which tracks predefined reference trajectories with rest of the members designated as followers which follow the leader or their neighbor [1,2,3,4]. Leader-following paradigm is easy to understand and implement. In addition, the formation can still be maintained even if the leader is perturbed by some disturbances. However, leader takes the most important role such that any failure of leader affects the whole system, and there is no feedback mechanism in this method, the error from one agent may be enlarged along the propagation.

The virtual structure approach is somewhat similar to the leader-following method, but the leader is virtual and entire formation is viewed as a rigid body.

* Mr.Ch. Venkatesh is Lecturer in Electrical Engg. Dept. at V.J.T.I, Sunil Surve is with Fr. Conceicao Rodrigues College of Engineering and he is pursuing as Research Scholar at V. J. T. I, Dr.N.M.Singh is Professor in Electrical Engg. Dept. at V.J.T.I.

This method maintains a rigid geometric relationship among each agent during movement; therefore it can't meet the requirements when the formation is time varying or need to change dynamically for some circumstances, such as obstacles avoidance, collision-free or deadlock-free paths.

The leader-following and virtual structure methods are more or less belongs to centralized control, however, most behavior-based methods is of the decentralized form. In behavior-based approach, several basic behaviors are prescribed for each agent, such as collision avoidance, obstacle avoidance, formation keeping, and goal seeking. Then final control is derived from weighted average of the control actions for each behavior. In this case, the changing of formation can be easily realized. Further more, some explicit feedback mechanism is included via communication network among the agents, which makes overall system more robust while avoiding the shortcoming of the leader-following control. However, formal stability analysis of the group behavior is not easy. Comparing with virtual structure approach, leader-follower paradigm can realize time-varying formation pattern. Even under complex conditions, such as uncertain parameters and unknown disturbances, individual control in the leader-follower paradigm can guarantee formation stability. Comparing with behavioral approaches, the feasibility of leader-following paradigm can be guaranteed mathematically. Therefore in this paper, we adopt leader-follower to realize formation in which information is directed, and robots are required to form formation according to certain formation pattern. The major problem is to develop a decentralized control strategy to guarantee formation's stability.

With reference to above background, the flatness-based feedback linearization has also received a lot of attention resulting in the earlier work reported by many researchers [5,6,7,8]. Although this technique has been applied to several nonlinear and linear mechanical systems [5,6,7,8]. The flatness based approach uses the characterization of system dynamics to generate a suitable output. In a situation where the output does not have a physical meaning or interpretation, the linearization could be done through a measurable system component that has a relationship to it.

Though flatness is intimately related to feedback linearization, in this paper a different viewpoint of flatness is followed, which uses a coordinate change based on a Lie-Backlund approach to equivalence, and flatness proposed by M. Fliess. The flat output provides the framework to derive the endogenous feedback compensator, which can result in a constant linear controllable system, for a given nonlinear system. The endogenous feedback linearization means existence of a compensator and a diffeomorphism, which preserves dimensions even after transformation. This control scheme retains the global stability due to a significance of flat output and the. In this paper an attempt has been made to highlight the formation of different vehicles using the flatness property. A method using flatness-based feedback linearization, controller has been proposed in this paper. The feedback linearization scheme requires the generation of a flat output from which the control law can be easily designed. The key contribution of paper is

to propose and develop a novel strategy of flatness-based feedback linearization for trajectory tracking and formation control of vehicles.

The proposed flatness-based control strategy has been validated using MATLAB simulation results for single and multiple vehicles. The simulation results have proved the effectiveness of the proposed control strategy, This paper is organised as follows: Section 2 discusses about the model of Non holonomic vehicle, section 3 emphasize on flatness-basics, section 4 emphasize on formation of vehicles, section 5 emphasize on flatness-based control strategy. Section 6 gives MATLAB simulation results and some observations and section 7 concludes the paper with future scope of work.

2 Non Holonomic Vehicle

This section presents a model [9] of a vehicle with 4 wheels rolling without slipping on the horizontal plane (O, X, Y) . We denote by (x, y) the coordinates of the point P, middle of the rear axle, Q the middle point of the front axle,

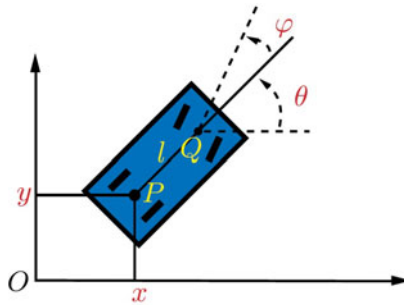


Fig. 1. Vehicle rolling without slipping on the plane

$\|\overrightarrow{PQ}\| = l$, θ the angle between the longitudinal axis of the vehicle and the Ox axis, and φ the angle of the front wheels (see Figure 1). The rolling without slipping condition, that geometrically corresponds to a specific kind of constraint called non holonomic constraint, which justifies the name of non holonomic vehicle given to this car idealization, reads: $\frac{d\overrightarrow{OP}}{dt}$ is parallel to \overrightarrow{PQ} and $\frac{d\overrightarrow{OQ}}{dt}$ is parallel to the frontwheels. Let us denote by $u = \|\frac{d\overrightarrow{OP}}{dt}\| \cdot \frac{\|\overrightarrow{PQ}\|}{\|\overrightarrow{PQ}\|}$ the modulus of the car’s speed. An elementary kinematic calculation yields:

$$\begin{aligned} \dot{x} &= u \cos \theta \\ \dot{y} &= u \sin \theta \\ \dot{\theta} &= \frac{u}{l} \tan \varphi \end{aligned} \tag{1}$$

3 Flatness-Basics

Flatness was first defined by Fliess et al. [5,8] using the formalism of differential algebra. In differential algebra, a system is viewed as a differential field generated by a set of variables (states and inputs). A model is described by a differential system. $\dot{x} = f_i(x, u)$ $i=1,2,3,\dots,n$

x_i denote the state variables and $u = (u_1, \dots, u_m)$ the control vector. The system is said to be flat if one can find a set of variables, called the flat outputs, such that the system is (non-differentially) algebraic over the differential field generated by the set of flat outputs. Roughly speaking, a system is flat if we can find a set of outputs (equal in number to the number of inputs) such that all states and inputs can be determined from these outputs without integration.

More recently, flatness has been defined in a more geometric context, where tools for nonlinear control are more commonly available. There are two different geometric frameworks for studying flatness and provide constructive methods for deciding the flatness of certain classes of nonlinear systems and for finding these flat outputs if they exist. One approach is to use exterior differential systems and regard a nonlinear control system as an affine system on an appropriate space [12]. In this context, flatness can be described in terms of the notion of absolute equivalence defined by E. Cartan [13]. Another geometric approach to study flatness is by using "Jet Bundles". In this paper a somewhat different geometric point of view is adopted, relying on a Lie-Backlund framework as the underlying mathematical structure. It offers a compact framework in which to describe basic results and is also closely related to the basic techniques that are used to compute the functions that are required to characterize the solutions of flat systems (the so-called flat outputs). In jet bundle approach a mapping from an infinite dimensional manifold whose coordinates are not only made up of original variables but also of jets of infinite order is dealt.

Fliess and coworkers have introduced the notion of an endogenous feedback which is essentially a dynamic feedback and they have shown that feedback linearisability via endogenous feedback is equivalent to differential flatness.

4 Formation Control of Vehicles

Moving in formation is a cooperative task and requires consent and collaboration of every agent in the formation. The formation problem can be defined as follows: Given desired formation for N mobile agents and a desired trajectory of the formation, the objective is for the agents to converge to the formation and to follow the desired trajectory avoiding obstacles while keeping the formation. To see this, consider a network of self-interested agents whose individual desire is to minimize their local cost $U_i(x) = \sum_{j \in N_i} \|x_i - x_j - r_{ij}\|^2$ via a distributed algorithm (x_i is the position of vehicle i with dynamics $\dot{x}_i = u_i$ and r_{ij} is a desired intervehicle relative-position vector). Instead, if the agents use gradient-descent algorithm on the collective cost $\sum_{i=1}^n U_i(x)$ using the following protocol:

$$\dot{x}_i = \sum_{j \in \mathbb{N}_i} (x_j - x_i - r_{ij}) = \sum_{j \in \mathbb{N}_i} (x_j - x_i) + b_i \tag{2}$$

with input bias $b_i = \sum_{j \in \mathbb{N}_i} r_{ij}$, the objective of every agent will be achieved.

5 Flatness Based Control Strategy

Flatness is undoubtedly related to the general problem of system equivalence. As a consequence, flatness is intimately related to feedback linearization. In this paper we propose a methodology based on Lie-Backlund approach to equivalence of systems. In this setting, two systems are said to be equivalent if any variable of one system may be expressed as a function of the variables of the other system and their finite number of time derivatives. Two such systems are then said to be isomorphic in the Lie-Backlund sense. Using this notion of Lie-Backlund isomorphism we show that the dynamics of Non Holonomic Vehicle corresponds to a trivial system by Endogenous feedback.

In implicit form, the system (1) becomes:

$$\dot{x} \sin \theta - \dot{y} \cos \theta = 0 \tag{3}$$

Let us show that this system is flat with (x, y) as flat output. The two first equations of [□](#) give

$$\tan \theta = \frac{\dot{y}}{\dot{x}}, u^2 = \dot{x}^2 + \dot{y}^2 \tag{4}$$

Differentiating the expression of $\tan \theta$, we get

$$\dot{\theta}(1 + \tan^2 \theta) = \frac{\ddot{y}\dot{x} - \dot{y}\ddot{x}}{\dot{x}^2}$$

from which we deduce

$$\dot{\theta} = \frac{\ddot{y}\dot{x} - \dot{y}\ddot{x}}{\dot{x}^2 + \dot{y}^2}$$

and, by the third equation of [□](#)

$$\tan \varphi = \frac{l\dot{\theta}}{v} = l \frac{\ddot{y}\dot{x} - \dot{y}\ddot{x}}{(\dot{x}^2 + \dot{y}^2)^{\frac{3}{2}}} \tag{5}$$

All the system variables x, y, θ, u, φ are thus expressed as functions of $x, \dot{x}, \ddot{x}, y, \dot{y}, \ddot{y}$

The system 1 is L-B equivalent to

$$\begin{aligned} \dot{x} &= v_1 \\ \ddot{y} &= v_2 \end{aligned}$$

This results in a constant linear controllable system, An elementary computation yields:

$$\begin{aligned} \ddot{x} &= \dot{u}\cos\theta - \frac{u^2}{l}\sin\theta\tan\varphi = v_1 \\ \ddot{y} &= \dot{u}\sin\theta + \frac{u^2}{l}\cos\theta\tan\varphi = v_2 \end{aligned}$$

and, after inversion of this linear system with respect to \dot{u} and $\tan\varphi$, we obtain the endogenous dynamic compensator:

$$\begin{aligned} \dot{u} &= v_1\cos\theta + v_2\sin\theta \\ \tan\varphi &= \frac{l}{u^2}(-v_1\sin\theta + v_2\cos\theta) \end{aligned}$$

If the whole state (x, y, θ) is measured, and if the speed u doesn't vanish, one can set, as before, the control law for the leader system(virtual)

$$v_1 = K_1(x - x_{ref}) - K_2 \int (x - x_{ref}) \tag{6}$$

$$v_2 = K_3(y - y_{ref}) - K_4 \int (y - y_{ref}) \tag{7}$$

with suitably chosen gains k'_i s, in order to ensure local exponential convergence of x, y and θ to their respective reference. The control law for the i^{th} followers is as given below:

$$v_1 = K_{i1} \sum_{j=1}^n a_{ij}(x_i - x_j) - K_{i2} \int (x_i - x_j) \tag{8}$$

$$v_2 = K_{i3} \sum_{j=1}^n a_{ij}(y_i - y_j) - K_{i4} \int (y_i - y_j) \tag{9}$$

These control schemes retain the global stability due to a significance of flat output . Thanks to the flatness property of the system that this is just an output tracker for the flat output (x, y) , since the state and input can be expressed in the coordinates $x, \dot{x}, \ddot{x}, y, \dot{y}, \ddot{y}$. For system represented in (1), MATLAB simulation has been carried out and the results have been represented here.

6 Simulation Results

We have used Leader-follower configuration for simulation. The adjacency matrix of the formation is given as:

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \text{ The Laplacian is } L = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} \text{ The controller (6, 7) is used}$$

for leaders while the controller (8, 9) is used for followers. We have used eight

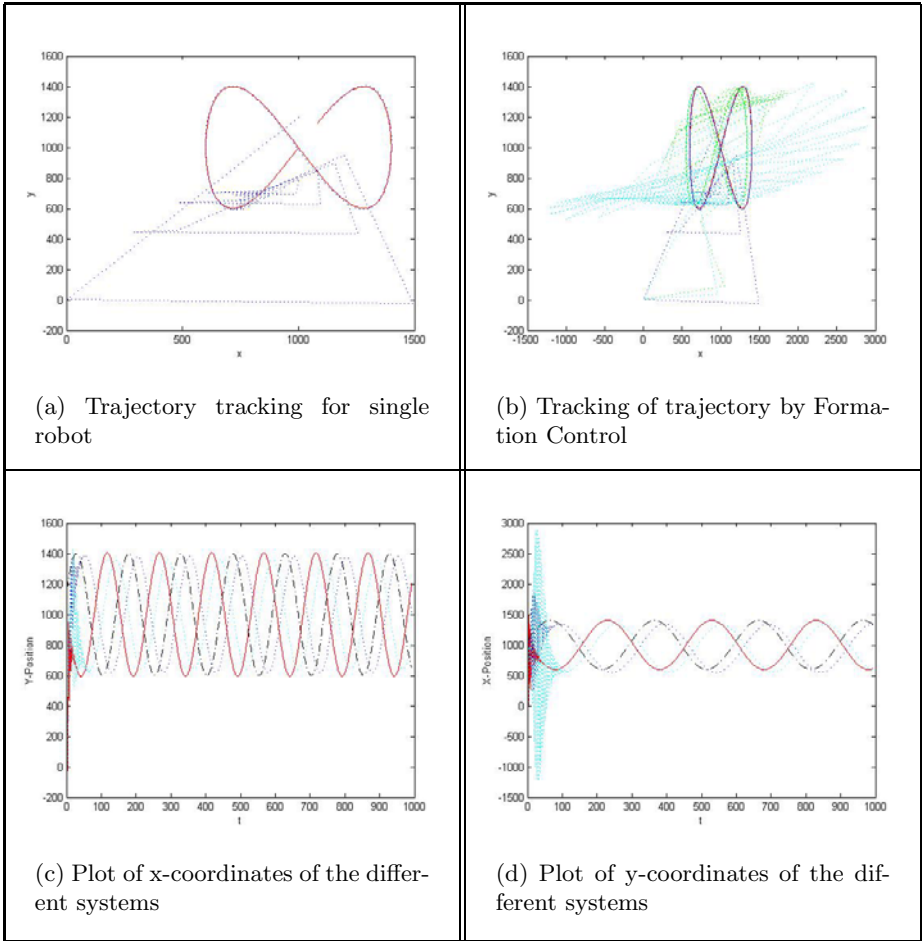


Fig. 2. Tracking of single and multiple robots

shape as reference trajectory as given below:

$$q_{ref}x = 1000 + 400 * \sin(2 * \pi * t(i)/30)$$

$$q_{ref}y = 1000 + 400 * \sin(4 * \pi * t(i)/30)$$

Figure 2(a) shows trajectory tracking for single robot. Figure 2(b) shows trajectory tracking for formation while figure 2(c) and figure 2(d) shows state variable plot different systems.

7 Conclusions

In this paper, we have proposed a novel scheme of formation controller using flatness. It is observed that the controller is very effective and ensures

guaranteed stability. The paper has given a different view point of flatness and have developed a flatness-based feedback linearization and shown its application in designing of formation controller which proved to be effective in tracking the desired trajectory. The future work is proposed to extend this idea of flatness scheme for virtual leader configuration.

References

1. Gu, D.: A Differential Game Approach to Formation Control. *The IEEE Transactions on Control Systems Technology* 16(1), 85–92 (2008)
2. Shao, J., Wang, L., Xie, G.: Flexible Formation Control for Obstacle Avoidance Based on Numerical Flow Field. In: *The Proceedings of the 45th IEEE Conference on Decision and Control*, San Diego, USA, December 13-15, pp. 5986–5991 (2006)
3. Sorensen, N., Ren, W.: A Unified Formation Control Scheme with a Single or Multiple Leaders. In: *The Proceedings of the 2007 American Control Conference*, New York, USA, July 11-13, pp. 5412–5418 (2007)
4. Hsu, H.C.-H., Liu, A.: Multiple Teams for Mobile Robot Formation Control. In: *The Proceedings of the 2004 IEEE International Symposium on Intelligent Control*, Taiwan, September 2-4, pp. 168–173 (2004)
5. Fliess, M., Levine, J., Martin, P., Rouchon, P.: Flatness and defect of non-linear systems: introductory theory and examples. *International Journal of Control* 61(6), 1327–1361 (1995) (online)
6. Levine, J., Nguyen, D.V.: Flat output characterization for linear systems using polynomial matrices. *Systems Control Letters* 48, 69–75 (2003)
7. Fliess, M., Levine, J., Martin, P., Rouchon, P.: A Lie-Backlund approach to equivalent and flatness of nonlinear systems. *IEEE Transactions on Automatic Control* 38, 700–716 (1999)
8. Martin, P., Murray, R.M., Rouchon, P.: Flat Systems Mini-Course ECC 1997 (1997)
9. Wang, J., Wu, X.-B., Xu, Z.-L.: Decentralized Formation Control and Obstacles Avoidance based on Potential Field Method. In: *The Proceedings of the Fifth International Conference on Machine Learning and Cybernetics*, Dalian, August 13-16, pp. 803–808 (2006)
10. Jia, Q., Li, G.: Formation Control and Obstacle Avoidance Algorithm of Multiple Autonomous Underwater Vehicles (AUVs) based on Potential Function and Behavior Rules. In: *The Proceedings of the IEEE International Conference on Automation and Logistics*, China, August 18 - 21, pp. 569–573 (2007)
11. Levine, J.: Analysis and control of non linear systems-A Flatness based approach. Springer, Heidelberg (2009)
12. van Nieuwstadt, M., Rathinam, M., Murray, R.M.: Differential flatness and absolute equivalence. In: *Proc. of the 33rd IEEE Conference on Decision and Control*, Lake Buena Vista, pp. 326–332 (1994)
13. Sluis, W.M.: Absolute Equivalence and its Application to Control Theory. Ph.D.thesis, University of Waterloo, Ontario (1992)

High Performance Tracking Controller for the Class of Uncertain Discrete-Time System with Input Delay

Deepti Khimani¹ and Machhindranath Patil²

¹ Vivekanand Education Society's Inst. of Tech.,
Mumbai, India, PIN-400071

² Indian Institute of Technology Bombay
Mumbai, India, PIN-400076

Abstract. In the general composite nonlinear feedback procedure the control input is assumed to be computed and applied at the same instant. However, the input delay adversely affects the performance of the system. In this paper the high performance controller design based on the predicted states for discrete-time input delay system with disturbance is presented and simulation of yaw angular displacement control of mini-helicopter is illustrated.

Keywords: Input delay system, Multirate output feedback, Composite nonlinear feedback.

1 Introduction

Time delay is commonly encountered in various electrical, hydraulic, mechanical systems and very often in chemical processes due to measurement, transmission and transportation lags, unmodelled inertia or computational delay. The existence of the delay often degrades the performance of the system and sometimes it leads to the instability. Input delay is caused by the transmission of a control signal over a long distance, state delay is a result of transmission or transport delay among interacting elements in a dynamic system and the output delay is the delay resulting from sensors.

Many different techniques are available in literature to solve the problem of stabilization of the time delay systems. A fundamental problem of eigen value assignment is proposed in [7], linear quadratic problem of the discrete-time lag system is addressed in [3]. A design procedure for the predictor based feedback controller for uncertain time-varying delay systems using linear matrix inequalities (LMI) is presented in [5]. In [1], [4], [20] predictor based sliding mode controller is designed for time delay systems.

In a state feedback control design the knowledge of complete state vector is essential but practically all the state variables are not available for measurement, nor it is suggested to measure them because of implementation complexities and increase in the cost. One of the effective ways to overcome this problem is

multirate output feedback. Multirate Output Feedback (MROF) is a discrete-time concept wherein either the control input or the sensor output is sampled at a faster rate than the other [13] - [20].

To improve the step response of second order linear systems with actuator saturation in terms of settling time and overshoot specifications, an idea of using CNF controller is proposed in [8] and further these results are extended to the higher order and multiple input systems by [9]. CNF controller consists of a linear feedback law and a nonlinear feedback law without any switching element. The linear feedback part is designed to yield a closed-loop system with a small damping ratio for a quick response and the nonlinear feedback law is used to increase the damping ratio of the closed-loop system as the system output approaches the target reference to reduce the overshoot caused by the linear part. See also [10], [11].

In this paper MROF based state predictor is derived and then high performance composite nonlinear feedback controller for predicted state is designed.

2 Problem Formulation

Consider the uncertain discrete time control system (Σ) with input delay

$$\Sigma : x(k + 1) = \Phi x(k) + \Gamma u(k - h) + Ed(k - h) \tag{1}$$

$$y(k) = Cx(k) \tag{2}$$

$$u(k) = \Theta(k) , k = -h, -h + 1, \dots 0 \tag{3}$$

Where, $x(k) \in \mathbb{R}^n$, $u(k) \in \mathbb{R}$, $y(k) \in \mathbb{R}$ are state, input and output of the system respectively. $d(k) \in \mathbb{R}$ is a disturbance, h is amount of delay and $\Theta(k)$ is an initial condition, which is generally available as it depends on the past inputs.

Assumptions

1. (Φ, Γ) pair is stabilizable.
2. (Φ, C) detectable and (Φ, Γ, C) has no invariant zero at $z=1$.
3. There exists the matrix F such that $(A + BF)$ is Hurwitz.

The assumption of stabilizability is imposed to facilitate the pole placement and the assumptions on the matrix C is made to allow output to track exogenous command input $r(k)$.

2.1 Discrete Time Predictor

In input delay system the state feedback control applied to the system at k^{th} sample comes into effect at $(k + h)^{th}$ sample. So the control law designed with states at k^{th} sampling period does not guarantee the performance and sometimes stability. However, this problem can be solved by predicting states at h samples

ahead and design the control law based on predicted states. The prediction of the state is possible by recursion. For the system (Σ) , we can write

$$\begin{aligned}
 x(k+2) &= \Phi x(k+1) + \Gamma u(k-h+1) + Ed(k-h+1) \\
 &= \Phi^2 x(k) + \Phi \Gamma u(k-h) + \Gamma u(k-h+1) \\
 &\quad + \Phi Ed(k-h) + Ed(k-h+1) \\
 x(k+3) &= \Phi^3 x(k) + \Phi^2 \Gamma u(k-h) + \Phi \Gamma u(k-h+1) \\
 &\quad + \Gamma u(k-h+2) + \Phi^2 Ed(k-h) \\
 &\quad + \Phi Ed(k-h+1) + Ed(k-h+2) \\
 &\quad \vdots \\
 x(k+h) &= \Phi^h x(k) + \Phi^{h-1} \Gamma u(k-h) \\
 &\quad + \Phi^{h-2} \Gamma u(k-h+1) + \Phi^{h-3} \Gamma u(k-h+2) \\
 &\quad + \dots + \Gamma u(k-1) \\
 &\quad + \Phi^{h-1} Ed(k-h) + \Phi^{h-2} Ed(k-h+1) \\
 &\quad + \Phi^{h-3} Ed(k-h+2) + \dots + Ed(k-1) \\
 \Rightarrow x(k+h) &= \Phi^h x(k) + \sum_{i=-h+1}^0 \Phi^{-i} \Gamma u(k+i-1) \\
 &\quad + \sum_{i=-h+1}^0 \Phi^{-i} Ed(k+i-1) \tag{4}
 \end{aligned}$$

Define the predicted state,

$$z(k) := \Phi^h x(k) + \sum_{i=-h+1}^0 \Phi^{-i} \Gamma u(k+i-1) + \sum_{i=-h+1}^0 \Phi^{-i} Ed(k+i-1) \tag{5}$$

Using the linear transformation (5), the original system (Σ) is converted into the the system with predicted states as follows.

$$\begin{aligned}
 z(k+1) &= \Phi^h x(k+1) + \sum_{i=-h+1}^0 \Phi^{-i} \Gamma u(k+i) + \sum_{i=-h+1}^0 \Phi^{-i} Ed(k+i) \\
 &= \Phi^h (\Phi x(k) + \Gamma u(k-h) + Ed(k-h)) + \sum_{i=-h+1}^0 \Phi^{-i} \Gamma u(k+i) \\
 &= \Phi^h \Phi x(k) + \Phi^h \Gamma u(k-h) + \Phi^h Ed(k-h) \\
 &\quad + \sum_{i=-h+1}^{-1} \Phi^{-i} \Gamma u(k+i) + \Gamma u(k) + \sum_{i=-h+1}^{-1} \Phi^{-i} Ed(k+i) + Ed(k) \\
 &= \Phi \Phi^h x(k) + \Phi \sum_{i=-h+1}^0 \Phi^{-i} \Gamma u(k+i-1)
 \end{aligned}$$

$$\begin{aligned}
 & + \Phi \sum_{i=-h+1}^0 \Phi^{-i} Ed(k+i-1) + \Gamma u(k) + Ed(k) \\
 & = \Phi \{ \Phi^h x(k) + \sum_{i=-h+1}^0 \Phi^{-i} \Gamma u(k+i-1) \\
 & \quad + \sum_{i=-h+1}^0 \Phi^{-i} Ed(k+i-1) \} + \Gamma u(k) + Ed(k) \\
 \Rightarrow z(k+1) & = \Phi z(k) + \Gamma u(k) + Ed(k) \tag{6}
 \end{aligned}$$

It is evident from the eq. (6) that $z(k)$ and $x(k)$ trajectories are generated from the same dynamical system. However, there is always time difference of h samples between these two trajectories. As the the original system is transformed into the delay-free system, the controller for the desired performance can be designed with $z(k)$ states.

2.2 Multirate Output Feedback (MROF)

As it is not always possible to measure states of the system, multirate output feedback (MROF) using fast sampled output can be used in place of a state observer to estimate the states of a discrete-time system. Also, the state feedback based control laws of any structure may be realized by the use of multirate output feedback [18].

The sampling period of the output is kept higher than or equal to the observability index of the system. Consider that the input $u(k)$ is applied with a sampling interval of T_s sec. and the system output is sampled at faster rate with sampling period of $\Delta = T_s/N$ sec, where N is an integer greater than or equal to the observability index.

Let $(\Phi_\Delta, \Gamma_\Delta, E_\Delta, C)$ represents the system sampled at Δ rate.

Define $v(k) := u(k-h)$, $\tilde{d}(k) := d(k-h)$ and rewrite the original system (Σ) with sampling period of T_s .

$$x((k+1)T_s) = \Phi x(kT_s) + \Gamma v(kT_s) + E\tilde{d}(kT_s) \tag{7}$$

$$y(kT_s) = Cx(kT_s) \tag{8}$$

Consider the Δ – system which is the original system sampled with Δ period.

$$\begin{aligned}
 x((k+1)\Delta) & = \Phi_\Delta x(k\Delta) + \Gamma_\Delta v(k\Delta) \\
 & \quad + E_\Delta \tilde{d}(k\Delta) \tag{9}
 \end{aligned}$$

$$y(k\Delta) = Cx(k\Delta) \tag{10}$$

Thus From eq. (9) we can write

$$x(kT_s + \Delta) = \Phi_\Delta x(kT_s) + \Gamma_\Delta v(kT_s) + E_\Delta \tilde{d}(kT_s) \tag{11}$$

As $v(k)$ and $\tilde{d}(k)$ are assumed to be unchanged during the interval $kT_s < t < (k + 1)T_s$, the $T_s - system$ dynamics can be constructed from the $\Delta - system$ dynamics. Using recursion procedure,

$$\begin{aligned}
 x(kT_s + \Delta) &= \Phi_\Delta x(kT_s) + \Gamma_\Delta v(kT_s) + E_\Delta \tilde{d}(kT_s) \\
 x(kT_s + 2\Delta) &= \Phi_\Delta x(kT_s + \Delta) + \Gamma_\Delta v(kT_s) \\
 &\quad + E_\Delta \tilde{d}(kT_s) \\
 &= \Phi_\Delta^2 x(kT_s) + (\Phi_\Delta + I)\Gamma_\Delta v(kT_s) \\
 &\quad + (\Phi_\Delta + I)E_\Delta \tilde{d}(kT_s) \\
 &\quad \vdots \\
 x(k + 1)T_s - \Delta &= \Phi_\Delta^{N-1} x(kT_s) + \sum_{i=0}^{N-2} \Phi_\Delta^i \Gamma_\Delta v(kT_s) \\
 &\quad + \sum_{i=0}^{N-2} \Phi_\Delta^i E_\Delta \tilde{d}(kT_s) \\
 x((k + 1)T_s) &= \Phi_\Delta^N x(kT_s) + \sum_{i=0}^{N-1} \Phi_\Delta^i \Gamma_\Delta v(kT_s) \\
 &\quad + \sum_{i=0}^{N-1} \Phi_\Delta^i E_\Delta \tilde{d}(kT_s) \tag{12}
 \end{aligned}$$

Comparing (7) and (12), it follows that

$$\Phi = \Phi_\Delta^N \tag{13}$$

$$\Gamma = \sum_{i=0}^{N-1} \Phi_\Delta^i \Gamma_\Delta \tag{14}$$

$$E = \sum_{i=0}^{N-1} \Phi_\Delta^i E_\Delta \tag{15}$$

Define N past output samples of multirate sampled system

$$Y_k := \begin{bmatrix} y(kT_s - T_s) \\ y(kT_s - T_s + \Delta) \\ \vdots \\ y(kT_s - \Delta) \end{bmatrix} \tag{16}$$

For simplicity we can write k for kT_s then $T_s - system$ with multirate sampled output can be represented as,

$$x(k + 1) = \Phi x(k) + \Gamma v(k) + E\tilde{d}(k) \tag{17}$$

$$Y_{k+1} = C_0 x(k) + D_0 v(k) + E_0 \tilde{d}(k) \tag{18}$$

Where

$$C_0 = [C_1 \ C_1\Phi_\Delta \ \dots \ C_1\Phi_\Delta^{N-1}]^T \tag{19}$$

$$D_0 = \left[0 \ C_1\Gamma_\Delta \ \dots \ C_1 \sum_{i=0}^{N-2} \Phi_\Delta^i \Gamma_\Delta \right]^T \tag{20}$$

$$E_0 = \left[0 \ C_1E_\Delta \ \dots \ C_1 \sum_{i=0}^{N-2} \Phi_\Delta^i E_\Delta \right]^T \tag{21}$$

2.3 State and Disturbance Estimation

As in [17], [18], the state can be estimated from past N multirate samples and immediate past input $v(k - 1)$.

$$x(k) = L_y Y_k + L_u v(k - 1) \tag{22}$$

The disturbance at the immediate past input is given by

$$\tilde{d}(k - 1) = M_2 Y_k - M_2 D_0 v(k - 1) \tag{23}$$

where

$$M = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix} = [C_0 \ E_0]^\dagger \tag{24}$$

$$L_y = \Phi M_1 + \Gamma M_2 \tag{25}$$

$$L_u = \Gamma - L_y D_0 \tag{26}$$

and $[\cdot]^\dagger$ is a general inverse of $[\cdot]$.

Substituting (22) into (5) we get the estimated predicted state,

$$\begin{aligned} z(k) &= \Phi^h L_y Y_k + \Phi^h L_u u(k - h - 1) \\ &+ \sum_{i=-h+1}^0 \Phi^{-i} \Gamma u(k + i - 1) \\ &+ \sum_{i=-h+1}^0 \Phi^{-i} E d(k + i - 1) \end{aligned} \tag{27}$$

Remark 1. Estimation of predicted state $z(k)$ is possible as the initial input sequence (3) is known. Therefore state feedback based control for system (6) can be designed with the estimated states $z(k)$ as in (27) with $\tilde{d}(k - 1)$ as in (23) is an estimate of $d(k)$.

3 High Performance Controller Design

A composite nonlinear feedback law that will make an output to track a command input speedily without experiencing overshoot is designed in three steps 1) linear feedback design 2) nonlinear feedback design and 3) Formulation of composite law by combining linear and nonlinear law.

3.1 Linear Feedback Design

There are several methods available in literature such as LQR, H_∞ and H_2 optimization approaches to design the state feedback law to achieve the desired transient specifications for the linear time invariant system. The state feedback control is designed by placing the poles for small rise time, however, it exhibits the overshoots as dominant poles of the closed-loop system $C(zI - \Phi - \Gamma F)^{-1}\Gamma$ has a small damping ratio. Classical design technique such as root locus for the single input-single output system can be used in designing F . For a tracking problem choose a linear feedback law for the system (6) without disturbance

$$z(k+1) = \Phi z(k) + \Gamma u(k) \quad (28)$$

$$u_L = Fz(k) + Gr(k) \quad (29)$$

Where $r(k)$ is a exogenous command input and G is a scalar that can be obtained as

$$G := [C(I - \Phi - \Gamma F)^{-1}\Gamma]^{-1} \quad (30)$$

Such a scalar G exists because by assumptions $(\Phi + \Gamma F)$ is stable and (Φ, Γ, C) is invertible and no invariant zeros on unit circle in complex z-plane.

Define $z_e(k) := z(k) - z_d(k)$

Where z_d is desired state computed as

$$z_d(k) := (I - \Phi - \Gamma F)^{-1}\Gamma Gr(k) \quad (31)$$

Following from eq. (29) and the definition of the $z_e(k)$

$$u_L = Fz_e(k) + Fz_d(k) + Gr(k) \quad (32)$$

From (28),

$$\begin{aligned} z(k+1) &= \Phi z_e(k) + \Phi z_d(k) + \Gamma u(k) \\ &= \Phi z_e(k) + \Phi z_d(k) + \Gamma Fz_e(k) \\ &\quad + \Gamma Fz_d(k) + \Gamma Gr(k) \\ &= (\Phi + \Gamma F)z_e(k) + (\Phi + \Gamma F)z_d(k) + \Gamma Gr(k) \end{aligned}$$

Substituting $Gr(k)$ from (31),

$$\begin{aligned} z(k+1) &= (\Phi + \Gamma F)z_e(k) + (\Phi + \Gamma F)z_d(k) \\ &\quad + (I - \Phi - \Gamma F)z_d(k) \\ z(k+1) &= (\Phi + \Gamma F)z_e(k) + z_d(k) \end{aligned} \quad (33)$$

As $r(k)$ is constant, $z_d(k+1) = z_d(k)$,

$$z_e(k+1) = (\Phi + \Gamma F)z_e(k) \quad (34)$$

By assumption $(\Phi + \Gamma F)$ is hurwitz, thus as $k \rightarrow \infty$ the error vector $z_e(k) \rightarrow 0$ and output $y(k) \rightarrow r$.

3.2 Nonlinear Feedback Design

The main purpose of adding the nonlinear part to the CNF controllers is to speed up the response by contributing a significant value to the control input when the tracking error $r - y$ is small. For faster response without exhibiting the overshoots initially the damping ratio of the closed loop poles is kept low and as it approaches the output the state feedback gain is changed to make the damping ratio high. Here the goal is to design a control law that varies the feedback gain without losing the stability. Let $P \in \mathbb{R}^{n \times n}$ be the positive definite matrix and $\rho(r, y)$ be the nonpositive function which is uniformly bounded and locally Lipschitz in y . Then nonlinear feedback control law is given by

$$u_N = \rho(r, y) \Gamma^T P \Phi_c z_e \tag{35}$$

Where $\Phi_c := (\Phi + \Gamma F)$. The matrix P can be obtained by solving the following Lyapunov equation for some positive definite matrix $W \in \mathbb{R}^{n \times n}$,

$$\Phi_c^T P + P \Phi_c = -W \tag{36}$$

The selection of nonlinear function is discussed in next section.

3.3 Composite Feedback Control Law

The linear and nonlinear feedback laws as in (29) and (35) are combined to form a composite nonlinear feedback controller

$$u(k) = u_L + u_N \tag{37}$$

$$\Rightarrow u(k) = Fz(k) + Gr(k) + \rho(r, y) \Gamma^T P z_e(k) \tag{38}$$

Thus for the tracking problem, the composite control law is given by

$$u(k) = (F + \rho(r, y) \Gamma^T P \Phi_c) z_e(k) \tag{39}$$

Where $\rho(r, y)$ is a nonlinear function satisfying the following condition.

$$2\rho(r, y) + \rho(r, y) \Gamma^T P \Gamma \rho(r, y) \leq 0 \tag{40}$$

The nonlinear function $\rho(r, y)$ is chosen such that its value changes from 0 to $-\beta$ that changes the initial gain F designed for small risetime to the gain K to yield overdamped response as tracking error tends to zero.

Possible choice of $\rho(r, y)$ is as follows,

$$\rho(y, r) = -\beta e^{-\alpha|y(k) - r(k)|} \tag{41}$$

Where β and α are tuning parameters. The value of β contributes the change in controller gain whereas α determines the rate of change of $\rho(r, y)$.

With the composite nonlinear control law the closed loop system without disturbance is represented as

$$\begin{aligned} z_e(k+1) &= (\Phi + \Gamma F) z_e(k) + \Gamma u_N \\ \Rightarrow z_e(k+1) &= \Phi_c z_e(k) + \Gamma \rho(r, y) \Gamma^T P \Phi_c z_e(k) \end{aligned} \tag{42}$$

Intuitively for the disturbance rejection, the control law (39) is modified as

$$u(k) = (F + \rho(r, y)\Gamma^T P\Phi_c)z_e(k) - \Gamma^\dagger Ed(k) \quad (43)$$

As the matched disturbance becomes unmatched in the discretization process and also because of varying rates, the predicted disturbance and actual disturbance are not matched. Let $\hat{d}(k)$ be the disturbance error. Hence the closed loop system with CNF law system with the disturbance can be represented as

$$z_e(k+1) = \Phi_c z_e(k) + \Gamma\rho(r, y)\Gamma^T P\Phi_c z_e(k) + \hat{d}(k) \quad (44)$$

3.4 Boundedness of the State $z(k)$

Consider the lyapunov function $V(k) = z^T(k)Pz(k)$. The forward difference $\Delta V = V(k+1) - V(k)$ along the solutions of the system (44) is given by

$$\begin{aligned} \Delta V(k) &= (z_e^T(k)\Phi_c^T + z_e^T(k)\Phi_c^T P\Gamma\rho(r, y)\Gamma^T)P(\Phi_c z_e(k) \\ &\quad + \Gamma\rho(r, y)\Gamma^T P\Phi_c z_e(k) - z_e^T(k)Pz_e(k) + \hat{d}^T(k)P\hat{d}(k)) \\ &= z_e^T(k)\Phi_c^T P\Phi_c z_e(k) + 2z_e^T(k)\Phi_c^T P\Gamma\rho(r, y)\Gamma^T P\Phi_c z_e \\ &\quad + z_e^T(k)\Phi_c^T P\Gamma\rho(r, y)\Gamma^T P\Gamma\rho(r, y)\Gamma^T P\Phi_c z_e \\ &\quad - z_e^T(k)Pz_e(k) + \hat{d}^T(k)P\hat{d}(k) \end{aligned} \quad (45)$$

Define $L := \Gamma^T P\Phi_c z_e$, substituting in eq. (45)

$$\begin{aligned} \Delta V(k) &= z_e^T(k)\{\Phi_c^T P\Phi_c - P\}z_e \\ &\quad + L^T\{2\rho(r, y) + \rho(r, y)\Gamma^T P\Gamma\rho(r, y)\}L \\ &\quad + \hat{d}^T(k)P\hat{d}(k) \end{aligned} \quad (46)$$

As $|\rho(r, y)| \leq \beta$ and from (36) and (40), it follows

$$\begin{aligned} \Delta V(k) &\leq -\lambda_{\min}(W)\|e(k)\|^2 - (2\beta + \lambda_{\min}(P))\|e(k)\|^2 \\ &\quad + d_m \end{aligned} \quad (47)$$

where d_m is bound for the disturbance term $\hat{d}(k)^T P\hat{d}(k)$. Thus $\Delta V(k) \leq 0$ if

$$\|e(k)\| \geq \sqrt{\frac{d_m}{2\beta + \lambda_{\min}(W) + \lambda_{\min}(P)}} := R \quad (48)$$

Hence the trajectory originated from outside of the ball of radius R with center at origin will enter the ball asymptotically.

3.5 Selection of P and β

To select the value of β classical root locus theory can be used. Converting the closed loop system into auxiliary system

$$G_{aux}(z) = C_{aux}(zI - \Phi_{aux})^{-1}\Gamma_{aux} \quad (49)$$

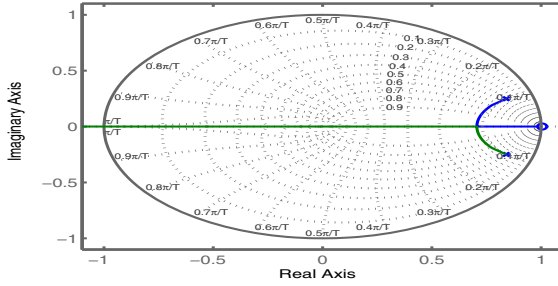


Fig. 1. Root locus of the auxiliary system

Where $\Phi_{aux} := \Phi_c$, $\Gamma_{aux} := \Gamma$ and $C_{aux} := \Gamma^T P \Phi_{aux}$. Note that G_{aux} is stable and invertible with $(n - 1)$ stable invariant zeros. The root locus starts from open loop poles i.e. eigen values of Φ_{aux} when the gain $\rho = 0$ and end up at open loop zeros including zero at infinity as ρ tends to infinity. So for the overdamped response at higher gains, part of the root locus should lie on real axis. The locations of invariant zeros depend on C_{aux} and thereby on P . This can be done by selecting appropriate W in Lyapunov equation to obtain $P > 0$. C_{aux} can also be obtained by the assignment of invariant zeros at desired locations on real axis [12]. Then P is determined from the definition of C_{aux} . The value of ρ at the desired root locations on locus is the absolute value of β .

4 Numerical Simulation

As an illustration yaw angular displacement control of mini-helicopter model as given in [2] is considered here. The discrete-time dynamics of the mini-helicopter with sampling period of $T_s = 0.08$ sec. is described by

$$\Phi = \begin{bmatrix} 0.7261 & 0 \\ 0.0685 & 1.0000 \end{bmatrix}, \Gamma = \begin{bmatrix} 1.0954 \\ 0.0461 \end{bmatrix}, E = \begin{bmatrix} 0.0014 \\ 0.0017 \end{bmatrix}, \text{ and } C = [0 \ 12.5] \quad (50)$$

The Δ -system with $N = 3$ is represented as

$$\Phi_{\Delta} = \begin{bmatrix} 0.8988 & 0 \\ 0.0253 & 1.0000 \end{bmatrix}, \Gamma_{\Delta} = \begin{bmatrix} 0.4047 \\ 0.0055 \end{bmatrix} \text{ and } E_{\Delta} = 10^{-3} \begin{bmatrix} 0.5059 \\ 0.5402 \end{bmatrix} \quad (51)$$

M_1 and M_2 for state and disturbance estimation using MROF as in section (2.3) is computed as follows

$$M_1 = \begin{bmatrix} -13.9643 & 14.0150 & 13.2629 & -13.3136 \\ 0.0764 & 0.0112 & -0.0116 & 0.0040 \end{bmatrix} \quad (52)$$

$$M_2 = [533.0332 \ -593.0332 \ -533.0332 \ 593.0332] \quad (53)$$

Choosing the state feedback matrix $F = [-0.0000 \ -0.9996]$ for the closed loop damping factor of 0.417.

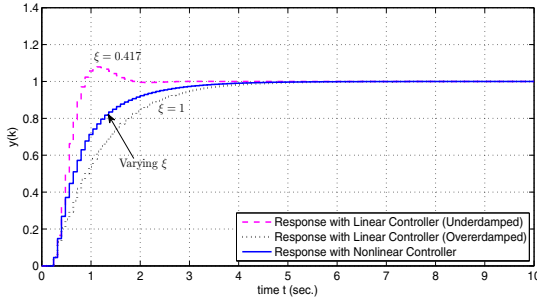


Fig. 2. The outputs of the system without disturbance using linear and composite nonlinear control laws

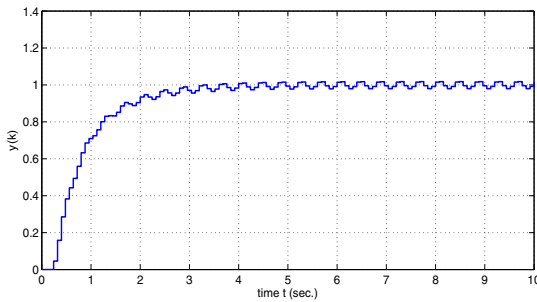


Fig. 3. The outputs of the system with disturbance using composite nonlinear control law

The nonlinear feedback controller parameters are chosen as,

$$\beta = 0.7 \text{ and } P = \begin{bmatrix} 1.0860 & 1.1941 \\ 1.1941 & 1.3276 \end{bmatrix} \text{ and tuning parameter } \alpha = 0.9.$$

The auxiliary system with chosen P is represented as

$$G(z) = \frac{1.096z - 1.095}{z^2 - 1.68z + 0.7676}$$

Fig. (1) shows the root locus of the auxiliary system.

The composite controller design results into the state feedback gain at the steady state, $K = [-0.7000 \quad -0.9996]$

In fig. (2) responses of linear control law with state feedback matrix F and K designed for underdamped and overdamped response are compared with the response of the system with composite nonlinear control law. Risetime of the response of the underdamped system is 0.8 sec and overdamped system is 2.4 sec with linear control laws. The composite nonlinear controller stabilizes the output in 9 sec with rise time of 1.8 sec.

The closed loop response of the system with disturbance $d(k) = 0.5\sin(5k)$ is shown in fig. (3). The control input of the composite nonlinear feedback controller

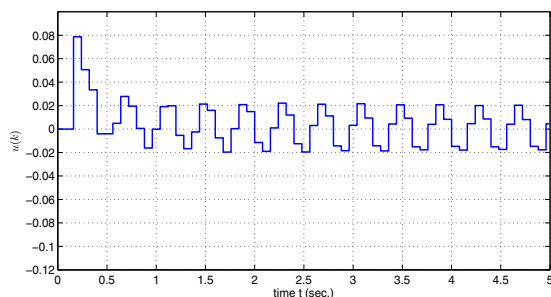


Fig. 4. The control input to the uncertain system

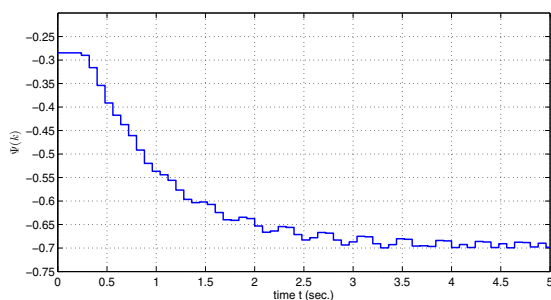


Fig. 5. Response of the nonlinear function $\rho(r, y)$

and the response of the $\rho(r, y)$ are shown in fig. (4) and fig. (5). The trajectories originated outside the ball of radius $R = 0.02$ are attracted toward the ball asymptotically.

References

1. Xia, Y., Liu, G.P., Shi, P., Rees, D., Liang, J.: A sliding mode control of uncertain linear discrete time systems with input delay. *IET Control Theory Appl.* 1(4), 1169–1175 (2007)
2. Lozanoa, R., Castilloa, P., Garciab, P., Dzulc, A.: Robust prediction-based control for unstable delay systems: Application to the yaw control of a mini-helicopter. *Automatica* 40, 603–612 (2004)
3. Pindyck, R.S.: The discrete-time tracking problem with a time delay in the control. *IEEE Transactions on Automatic Control* 17(3), 397–398 (1972)
4. Pang, H.-P., Liu, C.-J., Meng, X.-Z.: Sliding mode control for linear uncertain systems with input and state delays. In: *Proceedings of the IEEE International Conference on Information Acquisition*, Weihai, Shandong, China (2006)
5. Shen, J.C.: Designing stabilizing controllers and observers for uncertain linear systems with time-varying delay. *IEE Proceedings on Control Theory and Applications* 144(4), 331–333 (1997)

6. Furutani, E., Araki, M.: Robust stability of state-predictive and smith control systems for plants with a pure delay. *International Journal of Robust and Nonlinear Control* 8(18), 907–919 (1998)
7. Manitius, A.Z., Olbrot, A.W.: Finite spectrum assignment problem for systems with delays. *IEEE Transactions on Automatic Control* 24(4), 541–552 (1979)
8. Lin, Z., Pachter, M., Banda, S.: Towards improvement of tracking performance—nonlinear feedback for linear systems. *International Journal of Control* 70(1), 1–11 (1998)
9. Turner, M.C., Postlethwaite, I., Walker, D.J.: Nonlinear tracking control for multivariable constrained input linear systems. *International Journal of Control* 73, 1160–1172 (2000)
10. Chen, B.M., Lee, T.H., Peng, K., Venkataramanan, V.: Composite nonlinear feedback control for linear systems with input saturation: theory and an application. *IEEE Trans. on Automatic Control* 48(3), 427–439 (2003)
11. Venkataramanan, V., Peng, K., Chen, B.M., Lee, T.H.: Discrete-time composite nonlinear feedback control with an application in design of a hard disk drive servo system. *IEEE Trans. on Control Sys. Technology* 11(1), 16–23 (2003)
12. Chen, B.M., Zheng, D.Z.: Simultaneous finite and infinite zero assignments of linear systems. *Automatica* 31, 643–648 (1995)
13. Kranc, G.M.: Input-output analysis of multirate feedback systems. *IRE Trans. Automat. Contr.* 3(1), 21–28 (1957)
14. Jury, E.: A note on multirate sampled-data systems. *IRE Trans. Automat. Contr.* 12(3), 319–320 (1967)
15. Chammas, A.B., Leondes, C.T.: Pole assignment by piecewise constant output feedback. *Int. Journal of Control* 29(1), 31–38 (1979)
16. Hagiwara, T., Araki, M.: Design of a stable state feedback controller based on the multirate sampling of plant output. *IEEE Trans. on Auto. Control* 33(9), 812–819 (1988)
17. Janardhanan, S., Kariwala, V.: Multirate-Output-Feedback-Based LQ-Optimal Discrete-Time Sliding Mode Control. *IEEE Trans. on Auto. Control* 53(1), 367–373 (2008)
18. Bandyopadhyay, B., Janardhanan, S.: *Discrete Time Sliding Mode Control*. LNCIS, vol. 392. Springer, Heidelberg (2006)
19. Franklin, G., Powell, J.D., Workman, M.L.: *Digital Control of Dynamic Systems*, 3rd edn. Prentice Hall, Englewood Cliffs (2005)
20. Bandyopadhyay, B., Fulwani, D., Kim, K.S.: *Sliding mode control with novel sliding surfaces*. LNCIS, vol. 392. Springer, Heidelberg (2009)

Implementation of Controller Area Network (CAN) Bus (Building Automation)

S. Ashtekar Shweta¹, D. Patil Mukesh², and B. Nade Jagdish³

¹ Lecturer, Ramrao Adik Institute of Technology, Nerul, Navi-Mumbai
shweta_sa06@yahoo.com

² Assistant Professor, Ramrao Adik Institute of Technology, Nerul, Navi-Mumbai
mdpatal@iitb.ac.in

³ VDF's Institute of Technology, Latur
nade.jag@gmail.com

Abstract. The Controller Area Network (CAN) is an asynchronous serial CSMA/CD+AMP communication protocol for microcontrollers networks, supporting distributed real-time control (bit rate up to 1Mbps) with a very high level of security. CAN communication protocol is based on a distributed scheme, there is no central unit, allowing a direct data transfer between any two or more nodes without a master node mediation. A Building Automation System (BAS) is an example of a distributed control system. In this paper, the CAN bus lighting network is implemented with a reliable two wire control which is required for saving energy consumption or, creating precision lighting effects as a subunit of a BAS. Security systems is also interlocked to a building automation system to monitor the secure premises, control the operation of the overall system, and authorize legal entries as well as to trigger the alarm.

Keywords: CAN protocol, PIC Microcontroller, Building Automation System.

1 Introduction

As consumer electronics, computer peripherals, vehicles, automation and industrial applications add embedded functionality, demand is growing for inexpensive, fast and reliable communication media to serve these applications. Today more and more of the building blocks used in embedded system design are replacing parallel buses with serial buses .CAN serial bus is the most applicative technology in the field of automation, which can transfer data at the rate up to 1Mbps. It has its own unique advantage with its reliability, flexibility and real-time performance .Most significant features of CAN Bus include: Broadcast or Multicast system, Multimaster structure-distributed control, Prioritization of messages, Flexible configuration (“hot pluggable”), Message routing, Remote data request (RTR), Max. speed 1Mbps (40 m) and 50 kbps (1000m), Reliable through error detection and recovery algorithms, automatic retransmission of corrupted messages. Building Automation is an idea of using a control system to monitor and command the mechanical, lighting, security control, or fire alarm systems in a commercial building. The computerized, intelligent network functions to keep building temperature within a specified range, control

lighting, monitor performance of all systems, and send out alarm signals to maintenance engineers or administrators when failure occurs. There are many controls in a building that can be included in a Building Automation System. HVAC controls, lighting controls, electricity controls, hot water controls, fire controls, access controls, security/ surveillance, vehicle parking system, plumbing system, lifts and elevators, gardening system in a Building Automation System[4]. The Building Automation System can be programmed to manage these controls. Commercially available lighting systems like DALI (Digital Addressable Lighting Interface) has disadvantages like Maximum 64 nodes can be interfaced, Slow speed 200 bps, Master/Slave configuration, Non flexible. These limitation can be removed using CAN bus network.

The CAN communication protocol is a CSMA/CD+AMP (Carrier Sense Multiple Access/Collision Detection+Arbitration on Message Priority) protocol. Every node on the network must monitor the bus for a period of no activity before trying to send a message on the bus (Carrier Sense). Also, once this period of no activity occurs, every node on the bus has an equal opportunity to transmit a message (Multiple Access). If two nodes on the network start transmitting at the same time, the nodes will detect the 'collision' and take the appropriate action. Messages remain intact after arbitration is completed even if collisions are detected. All of this arbitration takes place without corruption or delay of the higher priority message.

CAN implements three layers of ISO/OSI reference model as physical layer, data link layer and application layer. The Hi-speed CAN physical layer is merely a twisted pair of wires with a 120 ohm termination resistor at each end and twisted wire drops to the individual CAN nodes. CAN Hi voltage with respect to ground changes between 2.5 to 4 volts nominal. CAN Lo changes from 2.5 to 1 volt. Therefore the difference between the two is either 0 volts (is logical "1") or 2 volts (is logical "0"). 0 is known as the "recessive" state and 2 volts is the "dominant" state. These two signals, CAN Hi and CAN Lo are 180 degrees out of phase. Bus idle is when the voltage difference is zero. At data link layer, CAN supports four different types of frames as data frame, error frame, remote frame and overload frame. Either 11 bit or 29 bit arbitration field to identify the message. In this application we are using 11 bit message identifier in the data frame.

Sequence of transmitting data on CAN bus

- Initially Bus is Idle
- All nodes will transmit data
- Highest priority node will get access of the bus and transmit the data, other nodes will enter into receiving mode
- Transmit node will wait for Acknowledgement
- If acknowledgement is received, send the next message otherwise wait and resend the same message again
- Send EOF (End Of Frame)bit and enter into receiving mode

Sequence of receiving data on CAN bus

- If frame is received without errors (by CRC calculations) send the acknowledgement,
- If frame is corrupted, wait for the corrected frame

- Send the received frame to Acceptance filter mechanism
- If frame is accepted-send it to FIFO memory of the controller otherwise discard the frame.
- Alert the host processor about the valid frame.

2 System Design: Hardware

The aim of the paper is to implement CAN protocol to control lighting network. Security systems is also interlocked to this network to monitor the secure premises, control the operation of the overall system, and authorize legal entries as well as to trigger the alarm[2].

The CAN protocol is been implemented using microcontroller based system and PC. The communication is done through only two wires. The system will be using two lighting nodes (two microcontroller systems), one interfacing unit node (microcontroller system) with computer and one alarm unit node. The bulbs and LDR sensors will be taken as node points in lighting network. The overall system is based on the integration several subsystems. All the nodes will be communicating with each other by sending different messages with the predefined identifiers. All the nodes present on the CAN bus will receive the same messages but with the frame filtering characteristics of the receiver, only matched identifiers will be accepted and then frame data will be received by the corresponding node. Depending on the different intensity values of the LDRs, relay and bulb on off condition will be changed and simultaneously display on each LCD display as well as computer. For security unit the password is set. Person entering the building must enter the correct password otherwise the entry will be restricted and triggering of alarm unit will be directed to main control unit. The system block diagram is as shown in figure 1.

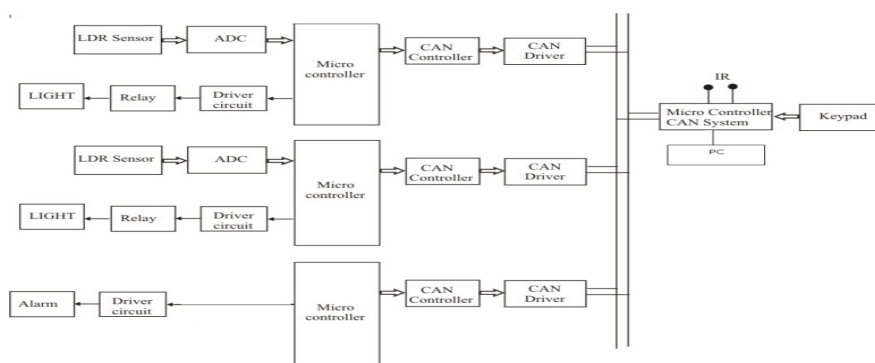


Fig. 1. Complete System Block Diagram for Building Automation using CAN Bus

Each of the nodes consists of a CAN Transceiver, CAN controller, PIC microcontroller. In a CAN bus system, each of the nodes are connected to the main node which has both the CAN Controller and CAN transceiver. CAN bus require only 2 wires (CANH and CANL) to connect the other nodes. CAN Transceiver MCP2551

plays a significant role in determining a successful data transmission over the can bus terminal. CAN transceiver is required to shift the voltage levels of the microcontroller to those appropriate for the CAN bus. This will help to create the differential signal CAN High and CAN Low which are needed in CAN bus.

MCP 2515 is standalone CAN controller has two acceptance masks and six acceptance filters that are used to filter out unwanted messages thereby reducing host MCUs overhead. The MCP 2515 interfaces with host MCU using industry standard SPI (Serial Peripheral Interface) as shown in figure 2 below. The main controller PIC16F877 includes the features like 10-bit, 8 channel ADC module, Synchronous Serial Port (SSP) with SPI and I2C, USART.

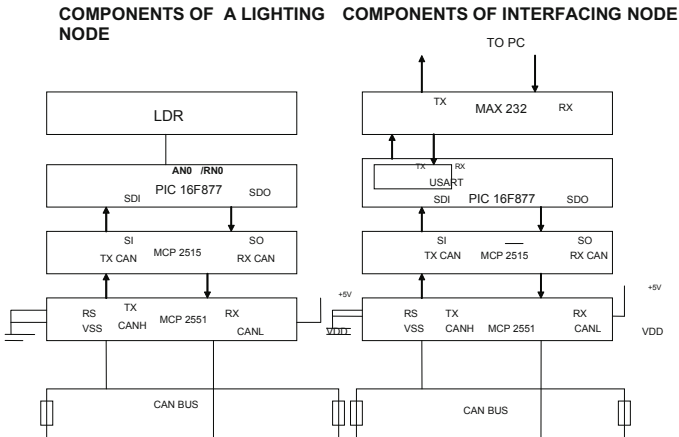


Fig. 2. Components of Lighting node and Interfacing node

Figure 2 shows different components used in Lighting and interfacing node. In the lighting nodes, the LDR sensors will measure the intensities of different values. The PIC microcontroller PIC16F877 will convert the data into digital form and through the SPI interface the data is sent to CAN controller MCP2515 and CAN transceiver. Finally data is sent to interfacing node which in turn will send the frame message to either turn on or off the relay. Lighting nodes will accordingly respond to this with message identifiers which are predefined and displays the required action on the LCD module. This interfacing node sends messages to both lighting nodes as well as to PC through USART interface. The converter MAX232 will convert CAN bus voltages to TTL logic levels [3]. It will transmit as well as receive different messages and controls overall system. In case of wrong authentication password, the error message will be forwarded to alarm unit and entry will be restricted. In the keypad node along with PIC controller and CAN controller, a keypad module is added using which the initial preset values are entered. Also the preset password is added for the security purpose. If the values exceeds or goes below preset values, the interfacing node will send the corresponding messages to respective nodes.

3 System Design: Software

The programming for PIC microcontroller is done by C- language by setting of different CAN controller registers and CAN configuration registers. The compiler used is Micro C compiler. The program is written for each of the individual nodes through which the message identifiers are set and to matched identifiers the respected node will respond .In case if the correct message is not r received the error message will be generated. The interface node main goal is to translate CAN 2.0A frames into serial port RS232 frames using MAX232 converter and vice versa. The system software allows bidirectional data transfer between nodes and PC through the interfacing node. The window to enter the display updated values of LDR intensities and to enter the password for authentication is developed using Visual Basics.

CANSPI Library routines:

The SPI module is available with a number of the PIC compliant MCUs. The mikroBasic PRO for PIC provides a library for working with CANSPI Add-on boards (with MCP2515 or MCP2510) via SPI interface. Some of them used in the project are,

CANSPIInitialize

```
sub procedure CANSPIInitialize(dim SJW as byte, dim BRP as byte, dim PHSEG1 as byte, dim PHSEG2 as byte, dim PROPSEG as byte, dim CANSPI_CONFIG_FLAGS as byte)
```

CANSPIRead

```
sub function CANSPIRead(dim byref id as longint, dim byref rd_data as byte[8], dim data_len as byte, dim CANSPI_RX_MSG_FLAGS as byte) as byte
```

CANSPIWrite

```
sub function CANSPIWrite(dim id as longint, dim byref wr_data as byte[8], dim data_len as byte, dim CANSPI_TX_MSG_FLAGS as byte) as byte
```

Lighting Node algorithm:

- Initialize I/O ports .Clear flags
- Configure different CANSPI modules.
- Read and convert LDR data
- Send the data to main control unit with TX_ID=2
- After delay, receive message with RX_ID=6
- Check first character of message and accordingly turn ON or OFF the relay

Main control unit algorithm:

- Set the limit value of nodes and password
- Process nodes value.
- For node 1, if process value>set value then turn on the relay1
- For node 2, if process value>set value then turn on the relay2
- Receive IR input
- Verify password. If yes ,process nodes value. If no ,turn the buzzer on.

4 Results and Analysis

A screen capture of the system with different lighting nodes and interfacing card is presented in figure 3. Complete hardware and software is tested. The desired results are verified using the designed system.

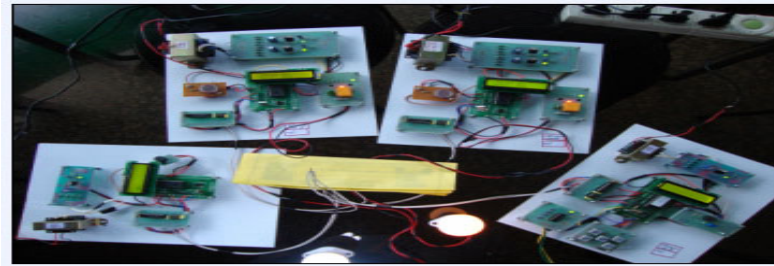


Fig. 3. Screen capture of system

A lighting node with two different intensity values displayed on the LCD and corresponding turning ON or OFF of the relays is represented in figure 4 and 5.

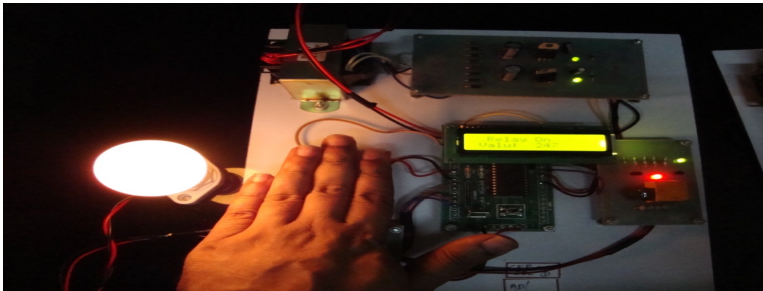


Fig. 4. Lighting node process value > set value, relay ON

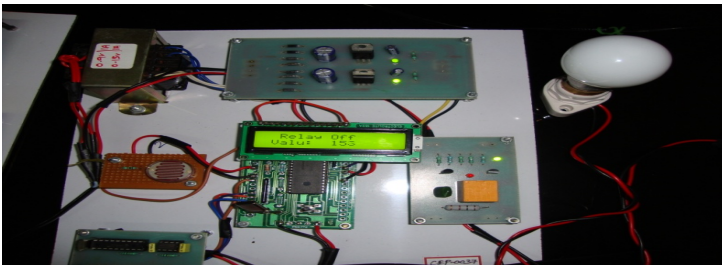


Fig. 5. Lighting node process value < set value, Relay OFF

A typical window developed in VB (Visual Basics) is as shown in figure 6 through which the user can enter the password for authentication .Also the displayed values of the lighting nodes are updated on this screen.

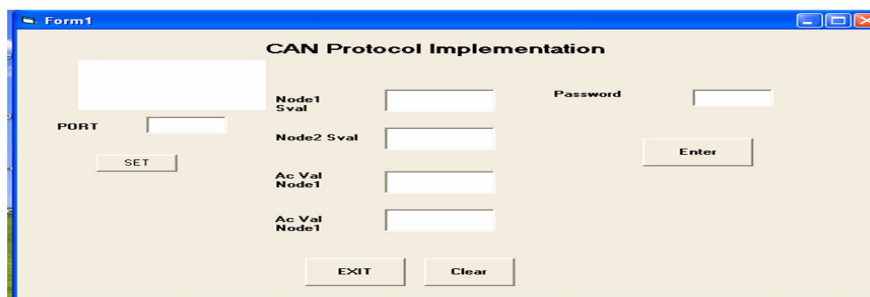


Fig. 6. GUI developed in VB for user interface

5 Conclusions

A small lighting network along with security provision allowing only authenticated users for building automation is implemented. It is composed of only four nodes, but the proposal can be expanded to a total number of 2048 network nodes and the use of any other kind of sensors required by individual subunits of BAS are possible[1]. This system requires only two wires and hence more efficient. The distributed processing from different component performed by different nodes reduces load on main controller thus increase in system performance. The proposed system based on microcontroller is found to be more compact, user friendly and less complex, which can readily be used in order to perform several difficult and repetitive tasks. The problems associated with wireless automation like physical obstructions, health concerns, data security, reliability, distance coverage can be overcome by implementing this CAN bus network.

The Controller Area Network (CAN) is an asynchronous serial CSMA/CD communication protocol for microcontrollers networks, supporting distributed real-time control (bit rate up to 1Mbps) with a very high level of security. Taking into account the different advantages of CAN bus a complete BAS (Building Automation System) for monitoring and controlling different subunits such as vehicle parking system[5], plumbing system, lifts and elevators, gardening system can be implemented.

References

- [1] Díaz, J., Rodríguez, E., Hurtado, L., Cacique, H., Ramírez, A., Vázquez, N.: LightNet a Reliable Option for Lighting Applications, enics. In: 2008 International Conference on Advances in Electronics and Micro-electronics, pp. 159–164 (2008)
- [2] Esro, M., Basari, A.A., Siva Kumar, S., Sadhiqin M I, A., Syariff, Z.: Controller Area Network (CAN) Application in Security System. World Academy of Science, Engineering and Technology 59 (2009)
- [3] Ran, P., Wang, B., Wang, W.: The Design of Communication Convertor Based on CAN Bus. In: Proceedings of the 2008 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, July 2 - 5 (2008)

- [4] Robles1, R.J., Kim1, T.-h.: Applications, Systems and Methods in Smart Home Technology: A Review. *International Journal of Advanced Science and Technology* 15 (February 2010)
- [5] Chou, L.-D., Sheu, C.-C., Chen, H.-W.: Design and Prototype Implementation of A Novel Automatic Vehicle Parking System. *International Journal of Smart Home* 1(1) (January 2007)
- [6] Dong, X., Wang, K., Zhao, K.: Design and Implementation of an Automatic Weighing System Based on CAN Bus. In: *Proceedings of the 2008 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, July 2 - 5 (2008)

DXCCII-Based Mixed-Mode Electronically Tunable Quadrature Oscillator with Grounded Capacitors

Mohd. Samar Ansari¹ and Sumit Sharma²

¹Dept. of Electronics Engineering, Aligarh Muslim University, Aligarh, India

²Z.H. College of Engg. & Tech., Aligarh Muslim University, Aligarh, India
mdsamar@gmail.com, sumitsharma@zhcet.ac.in

Abstract. A new mixed-mode quadrature oscillator circuit using two dual-X current conveyors (DXCCII)s and grounded passive components is presented. In the proposed circuit, two quadrature voltage-mode signals and a current-mode sinusoidal waveform are simultaneously available. The oscillation condition and oscillation frequency are independently controllable. Further, electronic tunability i.e. voltage control of the circuit is also discussed. The use of only grounded capacitors makes the proposed circuit suitable for integrated circuit implementation. Results of PSPICE simulation confirm the proposed theory.

Keywords: Dual-X current conveyor, mixed-mode circuit, quadrature oscillator, current-mode, voltage-mode, voltage controlled oscillator, electronic tunability.

1 Introduction

Sinusoidal oscillators constitute an important unit in many communication, power electronics and instrumentation systems. A special class of oscillators is one which is capable of generating outputs in phase quadrature. Such quadrature oscillators find applications in mixers and single-sideband generators in the field of communications, and for measurement purposes in vector generators in the case of instrumentation systems [1-7]. A repertoire of circuits for generating signals in phase quadrature is available in the technical literature. Some of them provided voltage-mode outputs [1, and the references therein] and others generated current-mode signals [2, and the references therein]. However, the recent popularity gained by mixed-mode circuits due to their versatility in analog signal processing applications has led to the development of quadrature oscillator circuits that can provide both voltage-mode and current-mode outputs simultaneously [3-7]. Other features of interest for such circuits are non-interactive control of the condition and frequency of oscillation, low component count and the use of grounded passive components from the viewpoint of monolithic integrated circuit implementation. For such circuits, the current controlled conveyor has been employed as the active building block of choice as it opens up the possibility of electronic tunability and resistor-less realizations [5-7]. However, not much work has been reported on mixed-mode quadrature oscillators using Dual-X current conveyors (DXCCII)s.

2 Existing Circuits

The technical literature is replete with mixed-mode quadrature oscillators [3-7]. The quadrature oscillator in [3] employs three resistors and two capacitors with a fully differential current conveyor (FDCC) to provide two voltage-mode and two current-mode outputs. Although the circuit of [3] uses only a single active element and exhibits non-interactive frequency control, electronic tunability is not present. The oscillator of [4] uses two CCCIs and two grounded capacitors to generate quadrature current-mode and phase-shifted voltage outputs. Electronic control of the frequency of oscillation as well as orthogonal control of frequency was available in [4]. However, the current outputs are not available at high impedance output nodes. The circuit of [5] provides two quadrature current outputs and two phase-shifted voltage outputs using one plus-type and two minus-type CCCIs but at the cost of the inclusion of a floating capacitor. The oscillator of [6] provides both voltage-mode and current-mode quadrature outputs but exhibits electronics control via an external current rather than a voltage. The oscillator circuit of [7] generates multiphase current outputs at high impedance, is electronically tunable and uses grounded capacitors. However, the circuit employs one translinear conveyor and three capacitors for each output of the n -phase oscillator.

In this paper, a versatile analog building block viz. the Dual-X Current Conveyor (DXCCII) is utilized to realize a CMOS compatible quadrature oscillator providing voltage-mode waveforms in phase quadrature. Another current-mode sinusoidal waveform is available simultaneously at a high-impedance node. Further, voltage-controlled resistors are used to replace the grounded resistances of the oscillator circuit to provide an electronic control on the frequency of the generated waveform. Results of computer simulations using PSPICE program are included to validate the operation of the proposed circuits.

3 Proposed Circuit

The Dual-X Current Conveyor (DXCCII) is a versatile analog building block for signal processing applications [8, 9]. Much of the versatility of DXCCII-based circuits is attributed to the presence of both the ‘normal’ and the ‘inverting’ X-terminals thereby resulting in two unique outputs at the Z+ and Z- terminals. This functionality essentially makes it a combination of a CCII and an ICCII. Fig. 1 shows the symbolic diagram of a typical DXCCII whose port relations are given in (1).

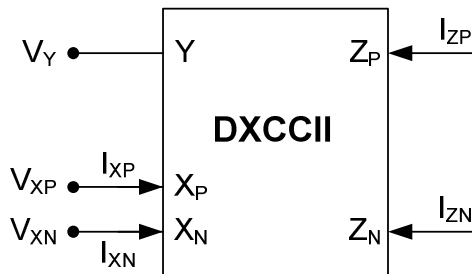


Fig. 1. DXCCII symbolic representation

$$\begin{bmatrix} I_Y \\ V_{XP} \\ V_{XN} \\ I_{ZP} \\ I_{ZN} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} V_Y \\ I_{XP} \\ I_{XN} \end{bmatrix} \quad (1)$$

Next, design of the proposed quadrature oscillator will be discussed. The basic scheme utilized here for the realization of a quadrature oscillator is presented in Fig. 2. The building blocks for this scheme, *viz.* an all-pass section and an integrator, are respectively characterized by the voltage transfer functions

$$T_1(s) = \frac{V_2}{V_1} = \frac{s - \alpha}{s + \beta} \quad (2)$$

$$T_2(s) = \frac{V_o}{V_2} = \frac{\beta}{s} \quad (3)$$

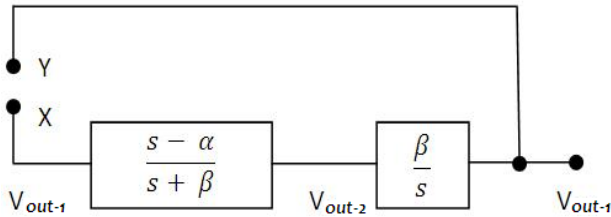


Fig. 2. Basic scheme to generate quadrature waveforms

where, α is the zero frequency for the all-passsection and β is the integrator's gain constant. The loop gain between the points X and Y can be expressed as:

$$\frac{V_o}{V_1} = \frac{s - \alpha}{s + \beta} \cdot \frac{\beta}{s} \quad (4)$$

This scheme can be set to provide oscillations, if the loop gain is unity for $s = j\omega$. Therefore,

$$\left. \frac{s - \alpha}{s + \beta} \cdot \frac{\beta}{s} \right|_{s=j\omega} = 1 \quad (5)$$

From the above equation, the condition and frequency of oscillations is obtained as:

$$\alpha = \beta \quad (6)$$

$$\omega = \sqrt{\alpha\beta} = \alpha \quad (7)$$

It can be shown that when equation (6) is satisfied, the voltage waveforms V_{out-1} and V_{out-2} are in phase quadrature. The phase difference between V_{out-1} and V_{out-2} is given by

$$\varphi = \pi - \tan^{-1} \frac{\omega}{\alpha} \tag{8}$$

Using equation (7) in (8) yields

$$\varphi = \pi - \tan^{-1}(1) = \frac{\pi}{2} \tag{9}$$

The circuit for voltage – controlled quadrature oscillator can be arrived at by replacing the all – pass section and the integrator of Fig.2 by their DXCCII based implementations, and is presented in Fig. 3. Using (6, 7), the frequency and condition of oscillation (FO and CO respectively) can be readily obtained as

$$FO: \omega_o = \frac{1}{\sqrt{R_2 R_3 C_2 C_3}} \tag{10}$$

$$CO: C_1 R_1 = C_3 R_3 \tag{11}$$

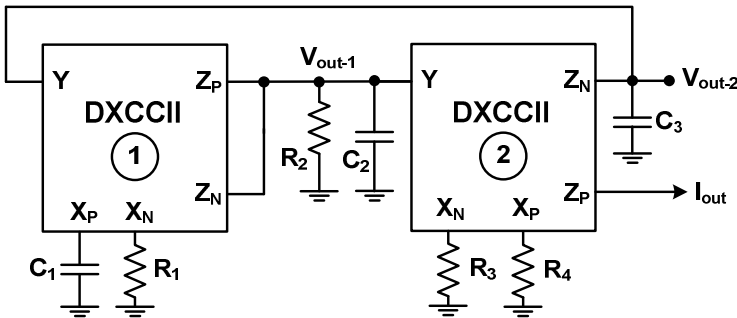


Fig. 3. Proposed DXCCII-based quadrature oscillator with grounded passive components

As is seen from equation (11), the condition of oscillation can be set by proper choice of the capacitors. Equation (10) may then be used to fix the frequency of oscillation independently by selecting suitable values of the resistors. Further, the proposed circuit is amenable from the viewpoint of VLSI fabrication in CMOS technology as all the passive components are grounded.

4 Results of PSPICE Simulations

The proposed quadrature oscillator was verified using PSPICE simulations carried out using an available CMOS implementation of DXCCII [9] in which 0.35 μm CMOS TSMC process parameters are utilized. The proposed circuit was simulated using $C_1 = C_2 = C_3 = 0.1nF$ and $R_1 = R_2 = R_3 = 1 K\Omega$ with C_1 adjusted to set the condition of

oscillation. The frequency of oscillation obtained is 1.29MHz (the design frequency being 1.59 MHz) and the time-domain view of the obtained quadrature voltage waveforms is shown in Fig.4 (a) and the frequency spectrum is plotted in Fig. 4 (b). It may be observed that the undesired harmonics are insignificant in comparison with the fundamental harmonic frequency of the circuit, thereby highlighting the high performance of the oscillator.

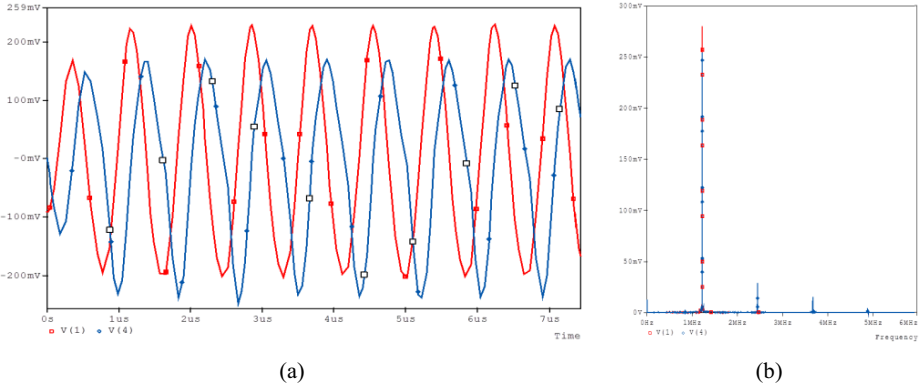


Fig. 4. (a): Simulation results for the proposed circuit showing quadrature voltage outputs. (b): Frequency domain representation of the quadrature voltage outputs.

The current-mode output I_{out} was also available simultaneously with the voltage-mode outputs. Time- and frequency-domain representations of the current-mode waveform are presented in Fig. 5 (a) and (b) respectively.

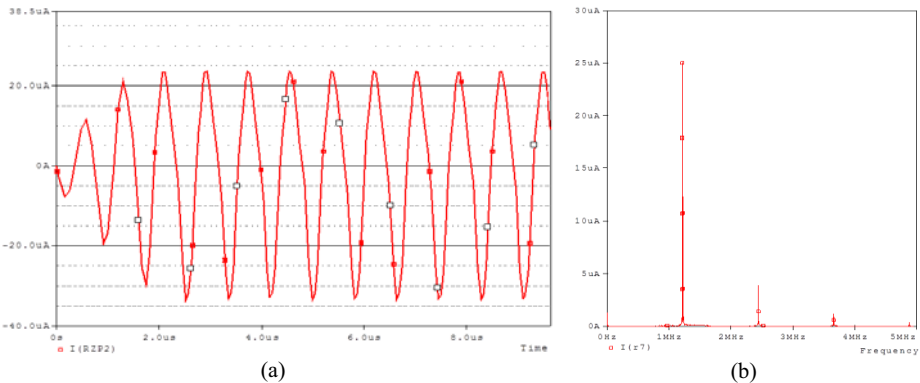


Fig. 5. (a): Simulation results for the proposed circuit showing the current-mode sinusoidal output. (b): Frequency domain representation of the current output.

Electronic control of the frequency of oscillation was explored by replacing the resistances R_1 , R_2 , and R_3 by voltage-controlled electronic resistances [10]. The

circuit is reproduced in Fig. 6 for reference. The variation in frequency with the control voltage (V_A) varying between 5V and 6V in steps of 0.5V is shown in Fig.7 (a) and the effect on the frequency spectrum is presented in Fig. 7 (b). Further, Fig. 7 (c) shows a plot depicting the variation of frequency with the control voltage which signifies a fairly linear control of frequency.

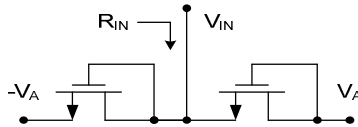


Fig. 6. Voltage-Controlled Electronic Resistance [10]

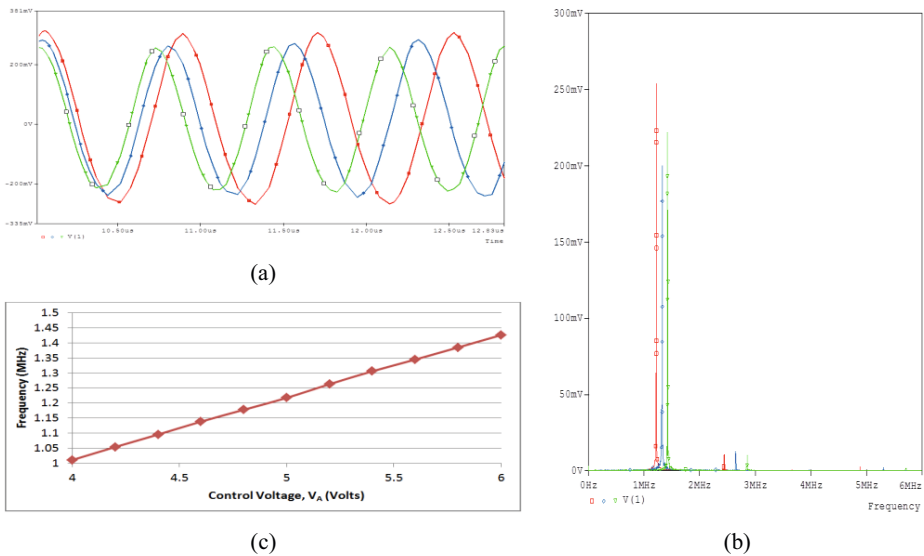


Fig. 7. (a): Simulation results showing variation in frequency of oscillation with control voltage. (b): Effect of control voltage on the frequency spectrum. (c): Frequency control by an externally applied control voltage.

5 Conclusion

A new mixed-mode quadrature oscillator employing only two DXCCII's and grounded resistors is proposed. The circuit is suitable for monolithic implementation by virtue of the use of grounded capacitors. The circuit provides two quadrature voltage outputs and a current-mode sinusoidal waveform. Electronic control of the frequency of oscillation was also demonstrated. The proposed circuit was verified using PSPICE.

References

- [1] Khan, I.A., Khwaja, S.: An integrable g^m -C quadrature oscillator. *Int. J. Electronics* 87(11), 1153–1157 (2000)
- [2] Biolek, D., Biolkova, V., Keskin, A.U.: Current mode quadrature oscillator using two CDTAs and two grounded capacitors. In: *Proc. 5th WSEAS Int. Conf. Sys. Sc. Sim. Engg.*, Spain, pp. 16–18 (December 2006)
- [3] Mohan, J., Maheshwari, S., Khan, I.A.: Mixed-Mode Quadrature Oscillators Using Single FDCCII. *J. of Active and Passive Electronic Devices* 2, 227–234 (2007)
- [4] Maheshwari, S.: Grounded Capacitor CM-APS with High Output Impedance. *Journal of Circuits, Systems and Computers* 16(4), 567–576 (2007)
- [5] Maheshwari, S.: New voltage and current-mode APS using current controlled conveyor. *Int. J. Electronics* 91(12), 735–743 (2004)
- [6] Ansari, M.S., Maheshwari, S.: Electronically tunable MOSFET-C mixed-mode quadrature oscillator. In: *Proc. IMPACT 2009, AMU, Aligarh (2009)*, Available on IEEExplore, doi:10.1109/MSPCT.2009.5164199
- [7] Abuelma'atti, M.T., Al-Qahtani, M.A.: A new current controlled multiphase sinusoidal oscillator using translinear conveyors. *IEEE Trans. CAS-II* 45, 881–885 (1998)
- [8] Zeki, A., Toker, A.: DXCCII-based tunable gyrator. *International Journal of Electronics and Communication (AEU)* 34(1), 59–62 (2005)
- [9] Minaei, S., Yuce, E.: A new full-wave rectifier circuit employing single dual-X current conveyor. *International Journal of Electronics* 95(8), 777–784 (2008), 1362-3060
- [10] Ibrahim, M.A., Minaei, S., Kuntman, H.: A 22.5 MHz current-mode KHN-biquad using differential voltage current conveyor and grounded passive elements. *Int. J. of Electronics and Communications* 59(5), 311–318 (2005)

Stereo Matching for 3D Building Reconstruction

Gaurav Gupta¹, R. Balasubramanian², M.S. Rawat¹,
R. Bhargava², and B. Gopala Krishna³

¹ Department of Mathematics, H. N. B. Garhwal University,
Srinagar, India

² Department of Mathematics, Indian Institute of Technology Roorkee, India

³ Space Applications Centre, Indian Space Research Organisation, Ahmedabad, India
guptagaurav.19821@gmail.com, balaiitr@ieee.org, hnbrawat@gmail.com,
rbharfma@iitr.ernet.in, bgk@sac.isro.gov

Abstract. In this paper, we present an approach for a 3D building reconstruction from stereo images using local matching techniques. The approach starts with local intensity based stereo matching techniques which utilizes, sum of square differences (SSD) and gradient-based matching techniques. The obtained disparities (left and right disparity) using these matching techniques are filtered out by applying cross-checking algorithm i.e. by comparing left-to-right and right-to-left disparity maps for increasing the reliability of the disparity map. After the aggregation process, the winner-take-all optimization is used to find the optimal disparity map. Furthermore, in order to obtain better disparity, a median filter is adopted for preserving boundary of image and effective removal of noise. The results show that the proposed scheme is reliable, accurate and robust to high resolution aerial images. Results are also compared with the ground data.

Keywords: Stereo matching, Sum of square difference, Gradient method, and aerial image.

1 Introduction

Reliable and accurate 3D reconstruction of buildings is important for many applications such as digital 3D city models etc. Large efforts are being directed towards the automation of building reconstruction because manual reconstruction of buildings from aerial images is time consuming and requires skilled personnel. The stereo matching techniques are used for automatically extracting the height of the buildings from aerial stereo images. The reconstruction of 3D buildings has a processing chain of many steps for the automatic extraction of 3D buildings directly from high-resolution stereo aerial images and the stereo matching techniques are playing a key role in it. Many authors have used different local stereo matching techniques such as normalized cross correlation; feature matching and area based matching [13, 14, 15] as well as global matching techniques to obtain the disparity map. Much work has been done on automatic

stereoscopic matching, and two distinct matching methods have emerged: feature based and area-based approaches [11, 12]. Feature-based matching consists of matching primitive sets extracted from each image. Common features in an urban environment are points of interest, segments, and linear structures [12]. The feature-based approach is appropriate for discontinuity, because depth discontinuities commonly appear as intensity discontinuities in the images. In area-based matching, each pixel matches from one image with their corresponding pixels in other image by measuring the similarity of grey-level value. The accuracy of feature-based approach relies on quality of edge segmentation.

Stereo matching is the most challenging problem in computer vision. The goal of stereo matching is to determine the disparity map that can be turned into depth map from two or more images taken from distinct view points. Stereo matching is an ill-posed problem. Hence the recovery of an accurate disparity map still remains challenging, generally due to poor texture regions, disparity discontinuous boundaries and occluded area. A broad overview on stereo matching can be found in [1, 2]. All the methods on stereo matching attempt to match pixels in one image with their corresponding pixels in the other image. These methods can be classified into local (window-based) and global methods. Local methods perform matching at each pixel, using intensity values within finite window whereas global methods incorporate explicit smoothness assumptions and determine disparity simultaneously by using various minimization techniques such as dynamic programming, intrinsic curves, graph cuts, belief propagation etc.

Local matching methods are very efficient to score matching in two images. In general, local matching methods can be classified into three broad categories: block matching, gradient-based methods and feature matching. Block matching methods seek to estimate disparity at a point in one image by comparing a small window about that point (the template) with a series of small regions extracted from the other image (the search region) [2]. The block matching method includes sum of squared differences (SSD) method, sum of absolute differences (SAD) method and a normalized cross coefficient method. Gradient-based methods seek to determine small local disparities between two images by formulating a differential equation relating motion and image brightness. This is robust to changes in camera gain and bias. The feature-based methods seek to find the pair wise corresponding features (two sets of features) between the reference and sensed images. The features can be edges, corners, end points, lines or curves, etc. Two classes of feature based matching are hierarchical feature matching and segmentation matching. These methods are also sensitive to the quality of the original segmentation.

In this paper, we present an approach for automatic 3D building reconstruction from aerial images. First, disparity map which results to depth map is obtained by applying two local methods, sum of square difference (SSD) and gradient method simultaneously. The sum of square difference (SSD) method seeks to estimate disparity at a point in one image by comparing a finite window about that point with a series of small regions extracted from the other image.

The sum of square difference (SSD) is computationally simpler than other block matching methods (Local Matching) such as normalized cross correlation (NCC) and sum of absolute difference (SAD) although SSD is sensitive to radiometric gain and bias. In order to eliminate sensitivity of radiometric gain and bias, we are using a gradient based matching method. Gradient based method seeks to determine small local disparities between two images by formulating a differential equation relating motion and image brightness [2]. A self-adapting dissimilarity measure with sum of absolute intensity differences (SAD) and a gradient based measure are used to calculate disparity in [4]. A cross-checking algorithm eliminating unreliable disparity estimation is used for increasing reliability of the disparity map. Cross-checking algorithm is very efficient to remove outlier as used in [3, 4]. After the aggregation process, the winner-take-all optimization is used to find the optimal disparity map. Many other authors have used the winner-take-all optimization to find the optimal disparity map [5, 6, 7]. Local winner-take-all optimization has also been used in real-time stereo applications [5, 8]. Few authors have shown that, with proper cost aggregation approaches, algorithms based on local WTA optimization performs better than many global optimization based techniques. Most local optimization methods still have difficulty to deal with the point ambiguity owing to insufficient or repetitive texture unlike WTA. A median filter is then implemented on the obtained disparity for preserving boundary of image and effective removal of noise. A median filter smoothes the data by replacing each data point with the median of the neighboring data points defined within a finite window. Finally, the depth information of buildings in the scene is calculated from the stereo pair by the obtained disparity that represents the displacement of corresponding pixels in the images.

2 Disparity Estimation

2.1 Stereo Matching

In this approach, the initial disparity maps are obtained by applying two local methods simultaneously, sum of square difference (SSD) and gradient. In SSD, for each pixel in the left image (reference image I_l), similarity scores are computed by comparing a finite, small window of size 3×3 centered on the pixel to a window in the right image (I_r), shifting along the corresponding horizontal scan line. The traditional sum-of-squared-differences (SSD) algorithm can be described as [1]:

1. The matching cost is the squared difference of intensity values at a given disparity.
2. Aggregation is done by summing up the matching cost over square windows with constant disparity.
3. Disparities are computed by selecting the minimal (winning) aggregated value at each pixel.

$$C_S(u, v, d) = \sum_{(u,v) \in W(u,v)} [I_l(u, v) - I_r(u + d, v)]^2 \quad (1)$$

where I_l and I_r represent the left and right images of stereo pair and d denotes the disparity at a point (u, v) in the right image.

One of the main drawbacks of SSD is its high sensitivity to radiometric gain and bias. In order to remove this we have used gradient based method which is insensitive to radiometric gain and bias. The disparity by gradient based methods is calculated as follows:

$$C_G(u, v, d) = \sum_{(u,v) \in W(u,v)} [\nabla_u I_l(u, v) - \nabla_u I_r(u + d, v)]^2 + \sum_{(u,v) \in W(u,v)} [\nabla_v I_l(u, v) - \nabla_v I_r(u + d, v)]^2 \quad (2)$$

where $W(u, v)$ represent the 3×3 surrounding window at position (u, v) .

We used the linear combination of SSD and Gradient method to obtain the final disparity map as follows:

$$E(u, v, d) = C_S(u, v, d) + \lambda * C_G(u, v, d), 0 \leq \lambda \leq 1 \quad (3)$$

where C_S and C_G are the disparities obtained by SSD and gradient measure respectively.

2.2 Disparity Enhancement

In this section, a cross-checking algorithm eliminating unreliable disparity estimation is used for increasing reliability of the disparity map. Let the pixel (u', v') in the matching image is corresponding to the pixel (u, v) in the reference image. The initial disparities are $d(u', v')$ and $d(u, v)$ in the matching and reference image respectively. If $d(u, v) \neq d(u', v')$, the pixel (u, v) is considered as an outlier. Here we are just comparing left-to-right and right-to-left disparity maps. The reliable correspondence is filtered out by applying a cross-checking test.

After the aggregation process, the winner-take-all optimization is used to find the optimal disparity map. Winner-take-all (WTA) algorithm takes the lowest (aggregated) matching cost as the selected disparity at each pixel whereas the other algorithms like dynamic programming (DP), graph cut (GC) etc require (in addition to matching cost) the smoothness cost. A limitation of this approach is that the uniqueness of matches is only enforced for one image (the reference image), while points in the other image might get matched to multiple points.

Furthermore, In order to obtain better disparity, a mode filter is adopted for preserving boundary of image and effective removal of noise. Median filter is applied to disparity obtained by cross checking test and WTA. In order to remove noise, there are several kinds of linear and nonlinear filtering techniques. Out of these, we adopted the median filter technique for preserving boundary of image and effective removal of noise. The window size used for filtering has been fixed to 7×7 .

3 Results and Discussions

In the given approach, first, two local matching techniques is used to obtain the disparity. Second, cross-check and WTA algorithm with a median filter is applied to increase the number of reliable disparity estimation. Third, winner-take-all optimization is used to find the optimal disparity. To illustrate the efficiency of the proposed matching algorithm, we used the test data set consists of three pair of stereo images: one 'tsukuba' [16] and two buildings of an urban area (aerial images) [10]. The ground resolution for the aerial images is 25 cm. All the stereo pair are already rectified. The tsukuba image is

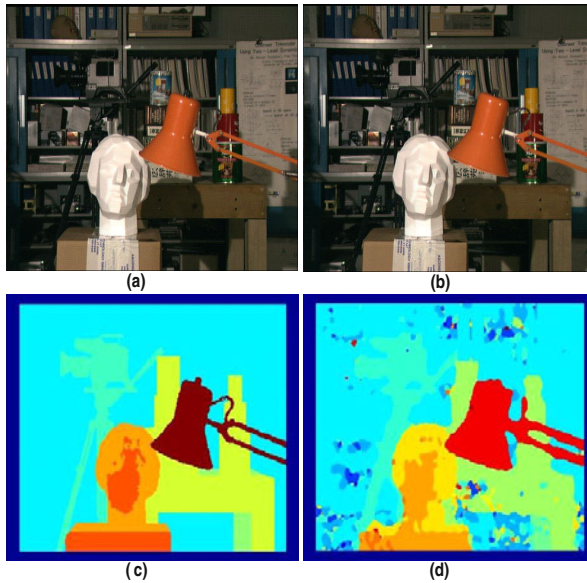


Fig. 1. The results on the Tsukuba data set. (a) and (b) stereo pair (c) Ground truth (d) Obtained segmentation result.

3.1 Error Analysis

First the obtained results (disparity) has been normalized in the range of ground truth. In case of tsukuba image, the error has been analyzed as the mean and standard deviation of depth error. The mean and standard deviation of absolute difference between the ground truth and obtained depth values has been calculated for the tsukuba data set. From the obtained results: mean 3.6 and standard deviation 5.2, it is concluded that the proposed algorithm is accurate enough for the reconstruction of 3D surfaces. The results for tsukuba data set is shown in figure 1. In case of aerial images, we selected positions (i.e., buildings)

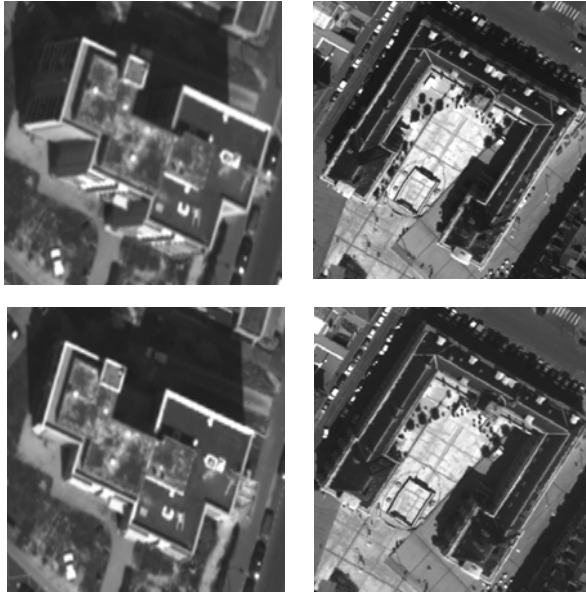


Fig. 2. Two stereo pair of high resolution aerial images

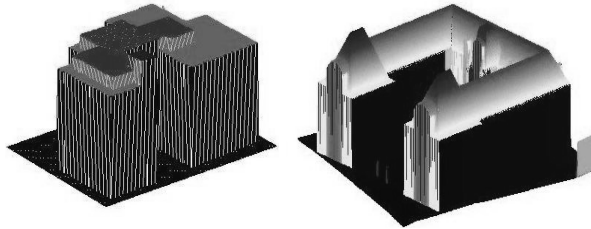
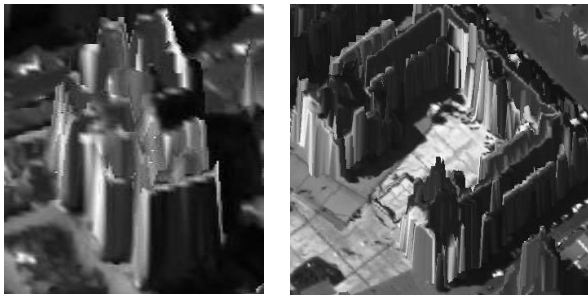


Fig. 3. True DSM (Digital Surface Model) of the buildings for above aerial images

in the interesting area in which we know the real values of those positions. Some sets of feature points are selected from each object (building). The selected sets of feature points from the reconstructed building roof are compared with the sets of feature points of original building (ground truth) roof. The experimental results getting from the proposed algorithm are shown in figure 4. For first building scene three sets of feature points are considered and the error for the scene is between 3 to 5 per-cent. For second building scene six sets of feature points are considered and the error for the scene is between 5 to 7 percent. Table 1 shows the results of applying our algorithm to two pairs of aerial images. The criteria for error analysis has used here however the availability of the given ground truth.

Table 1. Error results for independent set of feature points

Building	No. of feature Points	Height of the feature points		Error
		Reconstructed	Ground Truth	
Three sets from Building I	3530	481.53	501	19.47(03%)
	450	506.11	530	23.89(04.51%)
	7196	453.44	473	19.56 (04.14%)
Five sets from Building II	6007	467.97	495	27.03(05.46%)
	4055	481.53	511	29.47(05.77%)
	4161	517.28	558	40.72(07.29%)
	1555	555.93	596	40.07(06.72%)
	102	593.89	614	20.11(03.28%)

**Fig. 4.** Depth maps obtained from the stereo pair of aerial images

4 Conclusions

A new approach for automatic 3D building reconstruction from aerial images has been proposed. The conjunction of stereo matching, cross-checking test, WTA as well as median filter yields satisfactory results as demonstrated on the aerial images of the buildings as well as test image. Error analysis has also been done by comparing the obtained building heights with the original building heights (Ground truth). Results show that the satisfactory reconstruction is obtained by the proposed approach. In future work, we plan to extend the proposed method to automatic 3D building reconstruction from complex aerial images.

Acknowledgements. We gratefully acknowledge the financial support of Indian Space Research Organisation (ISRO), Ahmedabad, India through the Project scheme ISR-295-MTD to carry out this research work.

References

1. Scharstein, D., Szeliski, R., Zabih, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In: IEEE Workshop on Stereo and Multi-Baseline Vision, Kauai, Hawaii, pp. 131–140 (2001)
2. Brown, M.Z., Burschka, D., Hager, G.D.: Advances in Computational Stereo. IEEE Trans. on Pattern Analysis and Machine Intelligence 25(8), 993–1008 (2003)
3. Hong, L., Chen, G.: Segment-based stereo matching using graph cuts. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, D.C., vol. 1, pp. 74–81 (2004)
4. Klaus, A., Sormann, M., Karner, K.: Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure. In: Proceedings of the 18th International Conference on Pattern Recognition, Hong Kong, vol. 3, pp. 15–18 (2006)
5. Zhang, Y., Gong, M., Yang, Y.H.: Real-time multi-view stereo algorithm using adaptive weight Parzen window and local winner-take-all optimization. In: Fifth Canadian Conference on Computer and Robotic Vision, Windsor, pp. 113–120 (2008)
6. Zhang, Y., Gong, M., Yang, Y.H.: Local stereo matching with 3D adaptive cost aggregation for slanted surface modeling and sub-pixel accuracy. In: 19th International Conference on Pattern Recognition, Tampa, Florida, pp. 1–4 (2008)
7. Gu, Q., Zhou, J.: A novel similarity measure under Riemannian metric for stereo matching. In: The 33rd IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, Nevada, pp. 1073–1076 (2008)
8. Wang, L., Gong, M., Gong, M., Yang, R.: How Far Can We Go with Local Optimization in Real-Time Stereo Matching. In: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission, USA, pp. 129–136 (2006)
9. Wolf, P.R., Dewitt, B.A.: Elements of Photogrammetry with application in GIS, 3rd edn. McGraw Hill, New York (1999)
10. International Society for Photogrammetry and Remote Sensing, http://isprs.ign.fr/packages/packages_en.html
11. Wooo, D., Nguyena, Q., Nguyen, T.Q., Parka, Q., Jungb, Y.: Building detection and reconstruction from aerial images. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Part B3B, Beijing, China, vol. XXXVII (2008)
12. Baillard, C., Dissard, O.: A stereo matching algorithm for urban digital elevation models. Photogrammetric Engineering and Remote Sensing 66(9), 1119–1128 (2000)
13. Noronha, S., Nevatia, R.: Detection and Modeling of Buildings from Multiple Aerial Images. IEEE Trans. on Pattern Analysis and Machine Intelligence 23(5), 501–518 (2001)
14. Cornou, S., Dhome, M., Sayd, P., Cnrs, U., Blaise, U., Clermont-ferr, P.: Building Reconstruction from N Uncalibrated Views, <http://citeseerx.ist.psu.edu/viewdoc/summary?doi:10.1.1.4.7385>
15. Yom, J.-H., Lee, D.C., Kim, J.W., Lee, Y.W.: Automatic recovery of building heights from aerial digital images. In: IEEE Proc. of Geoscience and Remote Sensing Symposium, Anchorage, AK, vol. 7, pp. 4765–4768 (2004)
16. <http://vision.middlebury.edu/stereo/data/> (viewed on 12/08/09)

Fingerprint Identification Using Sectionized Walsh Transform of Row and Column Mean

H.B. Kekre¹, Tanuja K. Sarode², and Rekha Vig¹

¹ Mukesh Patel School of Technology and Management Engineering, Mumbai

² Thadomal Shahani College of Engineering, Mumbai

hbkekre@yahoo.com, tanuja_0123@yahoo.com, rekha.vig@nmims.edu

Abstract. Automated fingerprint identification systems (AFIS) are based on techniques to identify an unknown fingerprint with those present in the database. These techniques can be either localized (generally minutiae-based) or globalized (generally transform-based). The current day applications have to deal with large databases and fast processing and hence new techniques with less processing time need to be developed. In this paper we develop a new technique for fingerprint identification which reduces the processing time considerably as compared to standard existing techniques. This technique is in the frequency domain where coefficients (feature vectors) are generated using Walsh transform. Sectionization techniques are used to reduce the number of feature vectors. Proposed method is evaluated on standard database. The experimental results show that this algorithm could correctly identify fingerprints with accuracy more than 93% in case of larger number of sectors.

Keywords: Sectionization; Walsh Transform; frequency domain; fingerprint identification.

1 Introduction

Every person is believed to have distinct fingerprints, even identical twins that have similar DNA have different fingerprints. This is the reason why fingerprint identification is one of the most widely used and reliable biometric identification methods. There has been considerable research in the area of fingerprint identification, but it still is a very challenging area, the challenges being certain practical issues like worn-out and fake fingerprints etc., on one hand and ease and speed of processing on the other hand. Fingerprint is the pattern of ridges and valleys (furrows) on the surface of the finger [1][8]. There is an increased use of automated fingerprint based identification in both civilian and law-enforced organizations. Age-old method has been visual verification of minutiae by fingerprint experts and is based on minutiae detection. Minutiae, which mean minute details, are present at the local level. These basically refer to the discontinuities like ridge endings and ridge bifurcation points, and others

like ridge crossovers, islands etc. The fingerprint identification process includes locating the position, type and number of these minutiae [2]. Many automated processes involved in fingerprint identification try to use the same technique of minutiae extraction and matching. The minutiae-based matching techniques first extract the local minutiae from the thinned image or the grayscale image and then match their relative placement in a given fingerprint with the stored template [7]. The performance of minutiae based techniques rely on the accurate detection of minutiae points. Although the minutiae-based matching is widely used in fingerprint verification, it has problems in efficiently matching two fingerprint images containing different number of unregistered minutiae points and is easily affected by orientation and shift of fingerprint. Further, it does not utilize a significant portion of the rich discriminatory information available in the fingerprints. Hence they are less reliable, more complex and time consuming.

Instead of local level based identification, techniques which apply at global level have also been increasingly used[9]. They may achieve higher computational efficiency than minutiae based methods. In addition, they may be the only choice when the image quality of the given fingerprint is low. The ridge structure in a fingerprint can be viewed as an oriented textured pattern having a dominant spatial frequency (based on the repetitive pattern of the ridges) and orientation in local neighbourhood. Frequency domain analysis, which is commonly used in image processing [3][4][5], can be applied to extract the features of fingerprint in frequency domain which include the minutiae and ridge information fused together. The frequency based algorithms use different transforms to extract the information as feature vectors and compare them with those stored in the database. These feature vectors are much smaller in size than the entire fingerprint image and hence use less storage and computation. In this paper we are using Walsh transform to generate a set of feature vectors of a fingerprint. The Walsh coefficients are further treated to reduce the number of feature vectors. Results are represented in terms of average number of sample matches in the database and also in terms of the first match obtained. This technique can be used for implementation where speed of computation is of utmost importance.

The rest of the paper is structured as follows. Section 2 describes the Walsh transform technique. Section 3 explains the proposed sectionization method. Section 4 explains how two methods based on row-mean vector and column-mean vector have been fused to obtain better accuracy. The experimental results are given in section 5. Finally, conclusions are given in section 6.

2 Walsh Functions and Transform

The Walsh transform is a powerful tool of linear system analysis for discrete signals. Images are discrete functions of two dimensions, when acquired by digital acquisition devices. Hence Walsh transform can be aptly used on images to generate coefficients in frequency domain.

The Walsh functions are generated by Walsh generator as shown in the Fig. 1.

The Fig. 2 shows the Walsh functions for $N=8$. Here C_0 represents the coefficient of the DC component and S_n and C_n represent the SAL and the CAL (analogous to sine and cosine) coefficients of the n^{th} sequency (analogous to frequency) component.

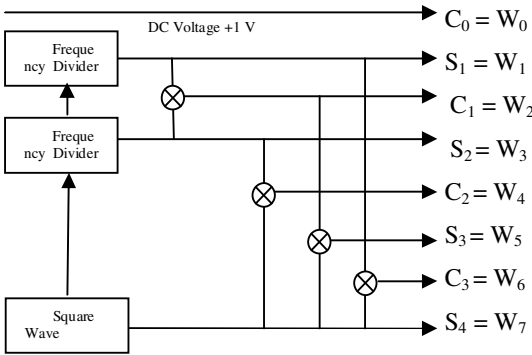


Fig. 1. Walsh Generator

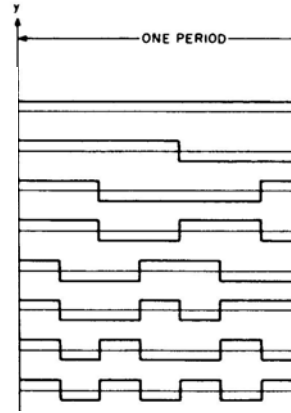


Fig. 2. Walsh Functions

The Walsh transform of one dimensional signal $f(x)$ can then be represented by the matrix as shown in equation (1) which is generated by sampling the Walsh function at the middle of the smallest time interval. The Walsh transform of a discrete signal $f(x)$ is calculated by (2)

$$W = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & j & j & j & -j & -j & -j & -j \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & j & -j & -j & j & j & -j & -j \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -j & -j & j & -j & j & j & -j \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -j & j & -j & j & -j & j & -j \end{pmatrix} \tag{1}$$

$$F(u) = W f' \tag{2}$$

where f' is the column representation of row vector form of discrete signal $f(x)$. The first value in $F(u)$ represents the DC component, and the next ones the higher sequency components. Now, since images are two dimensional the Walsh transform of an image is generated by two step method as shown in the Fig. 3. Here, first transform of each row vector is calculated and then the transform of each column vector of the output of first step is calculated to get the final Walsh transformed image.

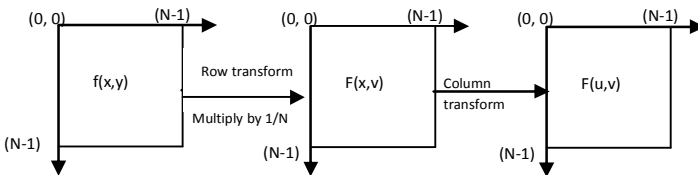


Fig. 3. Transform of a 2D function

Alternatively, the 2-D Walsh transform can be calculated by equation (3) where W and W' (transpose of Walsh matrix) are same, W being symmetric.

$$F(u,v) = (W f(x,y) W') / N \tag{3}$$

This two step method is very computation intensive and hence in this paper we have used a method to reduce the processing time. The Walsh transform is computed on two sets of vector: one is a row vector generated by calculating the mean of each column of the image matrix and the other is the column vector generated by calculating the mean of each row of the image matrix as shown in the Fig. 4. [6]

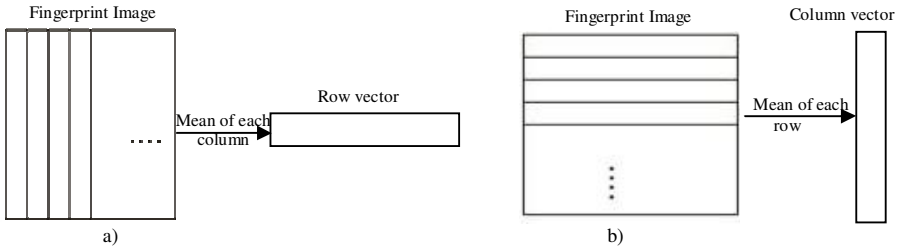


Fig. 4. a) Mean of columns b) Mean of rows of the fingerprint image

The Walsh transform of the row-mean vector is calculated by taking its transpose and that of the column-mean vector is calculated directly using equation (2) as the row mean and column vectors are now 1-D vectors. This method takes less processing time as row-mean and column-mean vectors are 1-D vectors and Walsh transform of 2 1-D vectors each of size N need $2N \log N$ additions whereas that of one 2-D vector (entire image) will need $2N^2 \log N$ additions.

3 Sectionization

The Walsh transforms of the fingerprint computed in two different manners, as discussed in previous section which generate a single column or single row matrix are subject to sectionization. Since the Walsh transform coefficients are integers, their absolute values are taken before further processing. The Walsh transform of row-mean vector and column-mean vector are divided into 4, 8, 12 or 16 sections. The mean of each section M_k is calculated as in equation (4), where W_i are the Walsh coefficients of the respective section. These are taken as feature vectors of that fingerprint image [10]. Hence the number of feature vectors generated is equal to the number of sections.

$$M_k = \frac{1}{N} \sum_{i=1}^N W_i(u) \tag{4}$$

These feature vectors are compared with the similar feature vectors of the database images and a similarity measure is generated. Here similarity measure used is Euclidean distance, which is calculated as sum of square of difference between the feature

vectors of test image and that of the database image. The database image whose Euclidean distance calculated is minimum is chosen as the best match.

4 Fusion

The results obtained from both the row and column mean methods are then fused together by using OR function to calculate the accuracy in terms of the first position matching and MAX function in terms of total number of matches obtained in the first 7 matches. Larger is the number of sectors better is the accuracy obtained in fingerprint identification.

5 Experimental Results

In this experiment, we have used the fingerprint image database containing 168 fingerprint images of size 256x256 pixels including 8 images per finger from 21 individuals. The set of fingerprints are obtained with rotational (+/- 15°) and shift (horizontal and vertical) variations as shown in Figure 5. The algorithm compared the feature vectors of the test image with those in database and the first 7 matched were recorded. The average number of matches in the first 7 matches gives a good indication of accuracy. Also whether the first match belonged to the same fingerprint or not was recorded. The results shown in Table 1 and 2 are of randomly selected 3 samples of each fingerprint and it has been observed that there is considerable improvement with increase in number of sectors as clearly shown in Fig. 6 and 7. We have applied the proposed technique on these images and results in each category show that our method can satisfactorily identify the fingerprint images. The accuracy rate observed is more than 93% for 16 sectors.

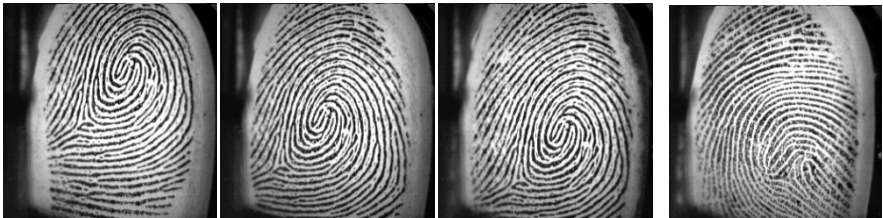


Fig. 5. Sample fingerprints of one finger

Table 1.

Sectors	Accuracy (in terms of first position match)
4	77.78 %
8	85.71 %
12	92.06 %
16	93.4%

Table 2.

Sectors	No. of Matches (Avg) in the first 7 matches
4	2.79
8	2.98
12	3.43
16	3.64

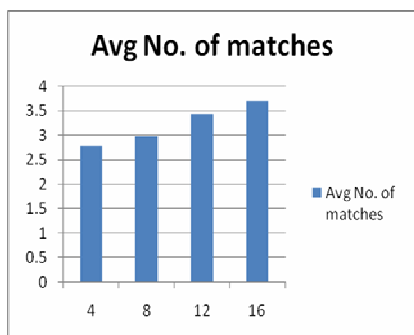


Fig. 6. Average number of matches in the first 8 matches for different sectors

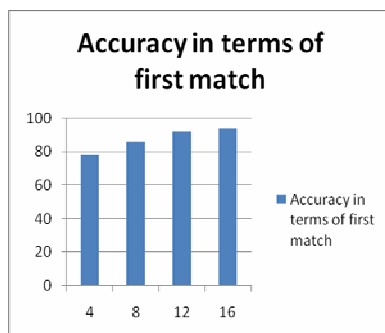


Fig. 7. Accuracy obtained for different sectors

6 Conclusion

In this paper, we have proposed a novel method for identification of fingerprint images. The technique of sectionization of the Walsh Transform of row and column means of fingerprint images has been used to generate feature vectors and matching is done using the same. As compared to other 1-D transform techniques like DCT used in [6], this method is computationally fast as it uses lesser number of features. It is also considerably independent of shift and rotation of fingerprint images. The results show that this simple yet robust method can be effectively used for fingerprint identification.

References

1. Maltoni, D., Maio, D., Jain, A., Prabhakar, S.: Handbook of Fingerprint Recognition. Springer, New York (2003)
2. Jain, L., et al.: Intelligent Biometric Techniques in Fingerprint and Face Recognition. CRC Press, Boca Raton (1999)
3. Gonzalez Rafael, C., Woods Richard, E.: Digital Image Processing, 3rd edn. Prentice Hall, Englewood Cliffs (2008)
4. Kekre, H.B., Sarode, T.K., Rawool, V.M.: Finger Print Identification using Discrete Sine Transform (DST). In: International Conference on Advanced Computing & Communication Technology (ICACCT-2008) Asia Pacific Institute of Information Technology, Panipat India, November 8-9 (2008)
5. Kekre, H.B., Sarode, T.K., Rawool, V.M.: Fingerprint Identification using Principle Component Analysis (PCA). In: International Conference on Computer Networks and Security (ICCNS 2008) held at VIT Pune, September 27-28 (2008)
6. Kekre, H.B., Sarode, T.K., Thepade, S.D.: DCT Applied to Column Mean and Row Mean Vectors of Image for Fingerprint Identification. In: International Conference on Computer Networks and Security (ICCNS 2008) held at VIT Pune, September 27-28 (2008)
7. Jain, A., Ross, A., Prabhakar, S.: Fingerprint matching using minutiae and texture features. In: Int'l Conference on Image Processing (ICIP), pp. 282–285 (October 2005)

8. Berry, J., Stoney, D.A.: The history and development of fingerprinting. In: Lee, H.C., Gaensslen, R.E. (eds.) *Advances in Fingerprint Technology*, 2nd edn., pp. 1–40. CRC Press, Florida (2001)
9. Ross, A., Jain, A., Reisman, J.: A hybrid fingerprint matcher. In: *Int'l Conference on Pattern Recognition (ICPR)* (August 2002)
10. Kekre, H.B., Mishra, D.: Four Walsh Transform Sectors Feature Vectors for Image Retrieval from Image Databases. Published in *International Journal of Computer Science and Information Technology (IJCSIT)* 01(02) (2010)

Author Index

- Acharya, Arup Abhinna 161
Ansari, Mohd. Samar 515
Anuradha, 342
Awale, R.N. 413
Awasthi, Ashish 1
Awasthi, Parul 200
- Baikerikar, Janhavi 116
Balasubramanian, R. 522
Banakar, R.M. 254
Banerjee, Joydeep 378, 393
Banerjee, Monalisa 349
Banerjee, Pradosh 349
Banerjee, Subhojit 393
Barpanda, Ravi Sankar 137
Barve, Amit 242
Bedi, Punam 25
Bhad, Sandeep 262
Bhadauria, Sarita Singh 200, 318
Bharambe, Ujwala 95
Bhargava, R. 522
Bhute, Avinash 457
Biswas, Debasish 335
Biswas, Sushanta 335
Biswas, Utpal 378, 393
- Chaki, Rituparna 275
Chakroborty, Indranil 378
Chand, Narottam 293
Chaudhari, R.S. 441
Chaudhuri, Atal 349
Chhabra, Jitender Kumar 448
Chunekar, Vaibhav N. 143
- Daga, Brijmohan 457
Dalal, Upena D. 433
Das, Sarita 355
Das, Soumik 349
DasGupta, Suparna 275
Dash, Subhasis 161
Deshmukh, A.M. 441
Dhanamma, Jagli 108
- Gawande, Kavita M. 79
Ghatol, Ashok 457
- Ghosh, Soumadip 335
Goel, Samiksha 328
Gopala Krishna, B. 522
Goutam, Aradhana 195
Gupta, Gaurav 522
- Haider, Mohammad 11
Halesh, M.R. 472
Hamed, Raed I. 49, 56
Hiwale, A.S. 262
Hota, Ashish Ranjan 18
- Ingle, Maya 195
- Jagdish, B. Nade 507
Jana, Tapas 378, 393
Javeri, Omkar 424
Jena, Sanjay Kumar 131, 300
Jeyakumar, Amutha 424
Jha, Rakesh Kumar 433
Jinaga, B.C. 254
- Kale, Achana 364
Kale, Archana 95
Karande, Aarti M. 143
Katiyar, Vivek 293
Katkar, Vijay 242
Kaushik, Awanish Kr. 342
Kekre, H.B. 530
Ket, Satish 413
Khare, Kavita 249
Khilar, Pabitra Mohan 400
Khimani, Deepti 494
Korra, Sathya Babu 131
Kothawale, Kausar R. 479
Kulkarni, Shrirang Ambaji 269
Kumar, Anubhav 231, 342
Kumar, Dinesh 448
Kumar, Jeet 1
Kumar, Mukesh 41, 150
Kumar, Vijay 448
- Lamgunde, Anuradha 364
Limkarl, Suresh 433

- Madhumita, Chatterjee 285
 Mahapatra, Jyoti Ranjan 300
 Mahapatro, Arunanshu 400
 Majhi, Banshidhar 137
 Mandar, Sohani 62
 Mavani, Monali 189
 Meshram, B.B. 143
 Mishra, Madhuri 200
 Mishra, Manoj Kumar 355
 Mitra, Souvik Kumar 393
 Mohanty, Jignyanshu 300
 Mukesh, D. Patil 479, 507
- Nag, Amitava 335
 Nand, Parma 406
 Naskar, M. K. 378, 393
 Nitnaware, Dhiraj 222
 Niyati, Marjan 306
- Panchal, V.K. 328
 Pandey, Mahendra Kumar 125
 Pandey, R.K. 231
 Pandya, H.N. 386
 Panigrahy, Saroj Kumar 131, 300
 Paradesi Rao, Ch.D.V. 465
 Parida, Bivasa Ranjan 355
 Parmar, Girish 125
 Pat, Ankit 18
 Patil, Machhindranath 494
 Patil, Sangita C. 79
 Patra, Tara Sankar 378
 Patsariya, Sanjay 125
 Prabhu, Sapna 116
- Raghavendra Rao, G. 269
 Rajan, E.G. 465
 Rajkamal 195
 Ramesh, Goparaju V. 70
 Ranjeet, K. 231
 Rao, Sattiraju N. 70
 Rao, Sudarshan 214
 Rasane, K.R. 472
 Rashmi, Salavi 62
 Rawat, M.S. 522
 Rinku, Shah 285
- Rohini, H. 472
 Rohini, Temkar 170
- Sahoo, Bibhudatta 137
 Sandeep, Chakravorty 88
 Sangita, Oswal 108
 Sarddar, Debabrata 378, 393
 Sarkar, Debasree 335
 Sarkar, Partha Pratim 335
 Sarode, Tanuja K. 530
 Sekhar Babu, S. 208
 Shabbiruddin, 88
 Shams Shafiqh, Alireza 306
 Sharma, Arpita 328
 Sharma, Mamta 189
 Sharma, Sumit C. 406, 515
 Sharma, Vishnu Kumar 318
 Shashi, Mogalla 70
 Shingate, Hemalata M. 214
 Shweta, S. Ashtekar 507
 Siddamal, Saroja V. 254
 Singh, Avneet 18
 Singh, Jaskirat 150
 Singh, N.M. 486
 Singhal, Archana 178
 Soni, Surender 293
 Sonia 178
 Sumalatha, V. 208
 Surve, Sunil K. 116, 486
- Tarasia, Nachiketa 355
 Thakur, Rakhi 249
 Thomas, Bindu A. 372
 Turuk, Ashok Kumar 137
- Unnikrishnan, Srija 214
- Vashisth, Pooja 25
 Venkatesh, Ch. 486
 Venkat Reddy, D. 465
 Venugopal, C.R. 372
 Verma, Ajay 222
 Vig, Rekha 530
 Vijay Kumar, T.V. 11
 Vyas, D.D. 386
- Yadav, R.L. 342