

Volume 3 Issue 11

November 2012



ISSN 2156-5570(Online)
ISSN 2158-107X(Print)



www.ijacsa.thesai.org



W H E R E W I S D O M S H A R E S

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS



THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org



Editorial Preface

From the Desk of Managing Editor...

IJACSA seems to have a cult following and was a humungous success during 2011. We at The Science and Information Organization are pleased to present the November 2012 Issue of IJACSA.

While it took the radio 38 years and the television a short 13 years, it took the World Wide Web only 4 years to reach 50 million users. This shows the richness of the pace at which the computer science moves. As 2012 progresses, we seem to be set for the rapid and intricate ramifications of new technology advancements.

With this issue we wish to reach out to a much larger number with an expectation that more and more researchers get interested in our mission of sharing wisdom. The Organization is committed to introduce to the research audience exactly what they are looking for and that is unique and novel. Guided by this mission, we continuously look for ways to collaborate with other educational institutions worldwide.

Well, as Steve Jobs once said, Innovation has nothing to do with how many R&D dollars you have, it's about the people you have. At IJACSA we believe in spreading the subject knowledge with effectiveness in all classes of audience. Nevertheless, the promise of increased engagement requires that we consider how this might be accomplished, delivering up-to-date and authoritative coverage of advanced computer science and applications.

Throughout our archives, new ideas and technologies have been welcomed, carefully critiqued, and discarded or accepted by qualified reviewers and associate editors. Our efforts to improve the quality of the articles published and expand their reach to the interested audience will continue, and these efforts will require critical minds and careful consideration to assess the quality, relevance, and readability of individual articles.

To summarise, the journal has offered its readership thought provoking theoretical, philosophical, and empirical ideas from some of the finest minds worldwide. We thank all our readers for their continued support and goodwill for IJACSA. We will keep you posted on updates about the new programmes launched in collaboration.

We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJACSA provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

We regularly conduct surveys and receive extensive feedback which we take very seriously. We beseech valuable suggestions of all our readers for improving our publication.

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 3 Issue 11 November 2012
ISSN 2156-5570(Online)
ISSN 2158-107X (Print)
©2012 The Science and Information (SAI) Organization

Editorial Board

Dr. Kohei Arai – Editor-in-Chief

Saga University

Domains of Research: Human-Computer Interaction, Networking, Information Retrievals, Optimization Theory, Modeling and Simulation, Satellite Remote Sensing, Computer Vision, Decision Making Methodology

Dr. Ka Lok Man

Xi'an Jiaotong-Liverpool University (XJTLU)

Domain of Research: Computer Science and Microelectronics

Dr. Sasan Adibi

Research In Motion (RIM)

Domain of Research: Security of wireless systems, Quality of Service

Dr. Zuqing Zuh

University of Science and Technology of China

Domains of Research : Optical Communication Systems, Optical network architecture and design, Next generation Internet, Signal processing, Broadband access network, such as cable access (DOCSIS) networks, passive optical networks (PON), fiber to the home (FTTH), Energy-efficient network and green technologies

Dr. Sikha Bagui

University of West Florida

Domain of Research: Database, database modeling, ER diagrams, XML data, web databases, data mining, association rule mining, data preprocessing

Dr. T. V. Prasad

Lingaya's University

Domain of Research: Bioinformatics, Natural Language Processing, Image Processing, Robotics, Knowledge Representation

Dr. Mohd Helmy Abd Wahab

Universiti Tun Hussein Onn Malaysia

Domain of Research: Data Mining, Database, Web-based Application, Mobile Computing

Reviewer Board Members

- **A Kathirvel**
Karpaga Vinayaka College of Engineering and Technology, India
- **A.V. Senthil Kumar**
Hindusthan College of Arts and Science
- **Abbas Karimi**
I.A.U_Arak Branch (Faculty Member) & Universiti Putra Malaysia
- **Abdel-Hameed A. Badawy**
University of Maryland
- **Abdul Wahid**
Gautam Buddha University
- **Abdul Hannan**
Vivekanand College
- **Abdul Khader Jilani Saudagar**
Al-Imam Muhammad Ibn Saud Islamic University
- **Abdur Rashid Khan**
Gomal University
- **Aderemi A. Atayero**
Covenant University
- **Ahmed Boutejdar**
- **Dr. Ahmed Nabih Zaki Rashed**
Menoufia University, Egypt
- **Ajantha Herath**
University of Fiji
- **Ahmed Sabah AL-Jumaili**
Ahlia University
- **Akbar Hossain**
- **Albert Alexander**
Kongu Engineering College, India
- **Prof. Alcinia Zita Sampaio**
Technical University of Lisbon
- **Amit Verma**
Rayat & Bahra Engineering College, India
- **Ammar Mohammed Ammar**
Department of Computer Science, University of Koblenz-Landau
- **Anand Nayyar**
KCL Institute of Management and Technology, Jalandhar
- **Anirban Sarkar**
National Institute of Technology, Durgapur, India
- **Arash Habibi Lashakri**
University Technology Malaysia (UTM), Malaysia
- **Aris Skander**
Constantine University
- **Ashraf Mohammed Iqbal**
Dalhousie University and Capital Health
- **Asoke Nath**
St. Xaviers College, India
- **Aung Kyaw Oo**
Defence Services Academy
- **B R SARATH KUMAR**
Lenora College of Engineering, India
- **Babatunde Opeoluwa Akinkunmi**
University of Ibadan
- **Badre Bossoufi**
University of Liege
- **Balakrushna Tripathy**
VIT University
- **Basil Hamed**
Islamic University of Gaza
- **Bharat Bhushan Agarwal**
I.F.T.M.UNIVERSITY
- **Bharti Waman Gawali**
Department of Computer Science & information
- **Bremananth Ramachandran**
School of EEE, Nanyang Technological University
- **Brij Gupta**
University of New Brunswick
- **Dr.C.Suresh Gnana Dhas**
Park College of Engineering and Technology, India
- **Mr. Chakresh kumar**
Manav Rachna International University, India
- **Chandra Mouli P.V.S.S.R**
VIT University, India
- **Chandrashekhara Meshram**
Chhattisgarh Swami Vivekananda Technical University
- **Chao Wang**
- **Chi-Hua Chen**
National Chiao-Tung University
- **Constantin POPESCU**
Department of Mathematics and Computer Science, University of Oradea
- **Prof. D. S. R. Murthy**
SNIST, India.
- **Dana PETCU**
West University of Timisoara
- **David Greenhalgh**
University of Strathclyde

- **Deepak Garg**
Thapar University.
- **Prof. Dhananjay R.Kalbande**
Sardar Patel Institute of Technology, India
- **Dhirendra Mishra**
SVKM's NMIMS University, India
- **Divya Prakash Shrivastava**
EL JABAL AL GARBI UNIVERSITY, ZAWIA
- **Dr.Dhananjay Kalbande**
- **Dragana Becejski-Vujaklija**
University of Belgrade, Faculty of organizational sciences
- **Driss EL OUADGHIRI**
- **Firkhan Ali Hamid Ali**
UTHM
- **Fokrul Alom Mazarbhuiya**
King Khalid University
- **Frank Ibikunle**
Covenant University
- **Fu-Chien Kao**
Da-Y eh University
- **G. Sreedhar**
Rashtriya Sanskrit University
- **Gaurav Kumar**
Manav Bharti University, Solan Himachal Pradesh
- **Ghalem Belalem**
University of Oran (Es Senia)
- **Gufran Ahmad Ansari**
Qassim University
- **Hadj Hamma Tadjine**
IAV GmbH
- **Hanumanthappa.J**
University of Mangalore, India
- **Hesham G. Ibrahim**
Chemical Engineering Department, Al-Mergheb University, Al-Khoms City
- **Dr. Himanshu Aggarwal**
Punjabi University, India
- **Huda K. AL-Jobori**
Ahlia University
- **Iwan Setyawan**
Satya Wacana Christian University
- **Dr. Jamaiah Haji Yahaya**
Northern University of Malaysia (UUM), Malaysia
- **Jasvir Singh**
Communication Signal Processing Research Lab
- **Jatinderkumar R. Saini**
S.P.College of Engineering, Gujarat
- **Prof. Joe-Sam Chou**
Nanhua University, Taiwan
- **Dr. Juan José Martínez Castillo**
Yacambu University, Venezuela
- **Dr. Jui-Pin Yang**
Shih Chien University, Taiwan
- **Jyoti Chaudhary**
high performance computing research lab
- **K Ramani**
K.S.Rangasamy College of Technology, Tiruchengode
- **K V.L.N.Acharyulu**
Bapatla Engineering college
- **K. PRASADH**
METS SCHOOL OF ENGINEERING
- **Ka Lok Man**
Xi'an Jiaotong-Liverpool University (XJTLU)
- **Dr. Kamal Shah**
St. Francis Institute of Technology, India
- **Kanak Saxena**
S.A.TECHNOLOGICAL INSTITUTE
- **Kashif Nisar**
Universiti Utara Malaysia
- **Kavya Naveen**
- **Kayhan Zrar Ghafoor**
University Technology Malaysia
- **Kodge B. G.**
S. V. College, India
- **Kohei Arai**
Saga University
- **Kunal Patel**
Ingenuity Systems, USA
- **Labib Francis Gergis**
Misr Academy for Engineering and Technology
- **Lai Khin Wee**
Technischen Universität Ilmenau, Germany
- **Latha Parthiban**
SSN College of Engineering, Kalavakkam
- **Lazar Stosic**
College for professional studies educators, Aleksinac
- **Mr. Lijian Sun**
Chinese Academy of Surveying and Mapping, China
- **Long Chen**
Qualcomm Incorporated
- **M.V.Raghavendra**
Swathi Institute of Technology & Sciences, India.
- **M. Tariq Banday**
University of Kashmir
- **Madjid Khalilian**

- Islamic Azad University
- **Mahesh Chandra**
B.I.T, India
 - **Mahmoud M. A. Abd Ellatif**
Mansoura University
 - **Manas deep**
Masters in Cyber Law & Information Security
 - **Manpreet Singh Manna**
SLIET University, Govt. of India
 - **Manuj Darbari**
BBD University
 - **Marcellin Julius NKENLIFACK**
University of Dschang
 - **Md. Masud Rana**
Khunla University of Engineering & Technology,
Bangladesh
 - **Md. Zia Ur Rahman**
Narasaraopeta Engg. College, Narasaraopeta
 - **Messaouda AZZOUZI**
Ziane AChour University of Djelfa
 - **Dr. Michael Watts**
University of Adelaide, Australia
 - **Milena Bogdanovic**
University of Nis, Teacher Training Faculty in
Vranje
 - **Miroslav Baca**
University of Zagreb, Faculty of organization and
informatics / Center for biomet
 - **Mohamed Ali Mahjoub**
Preparatory Institute of Engineer of Monastir
 - **Mohammad Talib**
University of Botswana, Gaborone
 - **Mohamed El-Sayed**
 - **Mohammad Yamin**
 - **Mohammad Ali Badamchizadeh**
University of Tabriz
 - **Mohammed Ali Hussain**
Sri Sai Madhavi Institute of Science &
Technology
 - **Mohd Helmy Abd Wahab**
Universiti Tun Hussein Onn Malaysia
 - **Mohd Nazri Ismail**
University of Kuala Lumpur (UniKL)
 - **Mona Elshinawy**
Howard University
 - **Monji Kherallah**
University of Sfax
 - **Mourad Amad**
Laboratory LAMOS, Bejaia University

- **Mueen Uddin**
Universiti Teknologi Malaysia UTM
- **Dr. Murugesan N**
Government Arts College (Autonomous), India
- **N Ch.Sriman Narayana Iyengar**
VIT University
- **Natarajan Subramanyam**
PES Institute of Technology
- **Neeraj Bhargava**
MDS University
- **Nitin S. Choubey**
Mukesh Patel School of Technology
Management & Eng
- **Noura Aknin**
Abdelamlek Essaadi
- **Om Sangwan**
- **Pankaj Gupta**
Microsoft Corporation
- **Paresh V Virparia**
Sardar Patel University
- **Dr. Poonam Garg**
Institute of Management Technology,
Ghaziabad
- **Prabhat K Mahanti**
UNIVERSITY OF NEW BRUNSWICK
- **Pradip Jawandhiya**
Jawaharlal Darda Institute of Engineering &
Techno
- **Rachid Saadane**
EE departement EHTP
- **Raghuraj Singh**
- **Raj Gaurang Tiwari**
AZAD Institute of Engineering and Technology
- **Rajesh Kumar**
National University of Singapore
- **Rajesh K Shukla**
Sagar Institute of Research & Technology-
Excellence, India
- **Dr. Rajiv Dharaskar**
GH Rasoni College of Engineering, India
- **Prof. Rakesh. L**
Vijetha Institute of Technology, India
- **Prof. Rashid Sheikh**
Acropolis Institute of Technology and Research,
India
- **Ravi Prakash**
University of Mumbai
- **Reshmy Krishnan**
Muscat College affiliated to stirling University.U
- **Rongrong Ji**
Columbia University
- **Ronny Mardiyanto**

- Institut Teknologi Sepuluh Nopember
- **Ruchika Malhotra**
Delhi Technological University
 - **Sachin Kumar Agrawal**
University of Limerick
 - **Dr.Sagarmay Deb**
University Lecturer, Central Queensland
University, Australia
 - **Said Ghoniemy**
Taif University
 - **Saleh Ali K. AlOmari**
Universiti Sains Malaysia
 - **Samarjeet Borah**
Dept. of CSE, Sikkim Manipal University
 - **Dr. Sana'a Wafa Al-Sayegh**
University College of Applied Sciences UCAS-
Palestine
 - **Santosh Kumar**
Graphic Era University, India
 - **Sasan Adibi**
Research In Motion (RIM)
 - **Saurabh Pal**
VBS Purvanchal University, Jaunpur
 - **Saurabh Dutta**
Dr. B. C. Roy Engineering College, Durgapur
 - **Sebastian Marius Rosu**
Special Telecommunications Service
 - **Sergio Andre Ferreira**
Portuguese Catholic University
 - **Seyed Hamidreza Mohades Kasaei**
University of Isfahan
 - **Shahanawaj Ahamad**
The University of Al-Kharj
 - **Shaidah Jusoh**
University of West Florida
 - **Shriram Vasudevan**
 - **Sikha Bagui**
Zarqa University
 - **Sivakumar Poruran**
SKP ENGINEERING COLLEGE
 - **Slim BEN SAOUD**
 - **Dr. Smita Rajpal**
ITM University
 - **Suhas J Manangi**
Microsoft
 - **SUKUMAR SENTHILKUMAR**
Universiti Sains Malaysia
 - **Sumazly Sulaiman**
Institute of Space Science (ANGKASA), Universiti
Kebangsaan Malaysia
 - **Sumit Goyal**
 - **Sunil Taneja**
Smt. Aruna Asaf Ali Government Post Graduate
College, India
 - **Dr. Suresh Sankaranarayanan**
University of West Indies, Kingston, Jamaica
 - **T C. Manjunath**
HKBK College of Engg
 - **T C.Manjunath**
Visvesvaraya Tech. University
 - **T V Narayana Rao**
Hyderabad Institute of Technology and
Management
 - **T. V. Prasad**
Lingaya's University
 - **Taiwo Ayodele**
Lingaya's University
 - **Tarek Gharib**
 - **Totok R. Biyanto**
Infonetmedia/University of Portsmouth
 - **Varun Kumar**
Institute of Technology and Management, India
 - **Vellanki Uma Kanta Sastry**
SreeNidhi Institute of Science and Technology
(SNIST), Hyderabad, India.
 - **Venkatesh Jaganathan**
 - **Vijay Harishchandra**
 - **Vinayak Bairagi**
Sinhgad Academy of engineering, India
 - **Vishal Bhatnagar**
AIACT&R, Govt. of NCT of Delhi
 - **Vitus S.W. Lam**
The University of Hong Kong
 - **Vuda Sreenivasarao**
St.Mary's college of Engineering & Technology,
Hyderabad, India
 - **Wei Wei**
 - **Wichian Sittipraporn**
Mahasarakham University
 - **Xiaoqing Xiang**
AT&T Labs
 - **Y Srinivas**
GITAM University
 - **Yilun Shang**
University of Texas at San Antonio
 - **Mr.Zhao Zhang**
City University of Hong Kong, Kowloon, Hong
Kong
 - **Zhixin Chen**
ILX Lightwave Corporation
 - **Zuqing Zhu**
University of Science and Technology of China

CONTENTS

Paper 1: Predicting Garden Path Sentences Based on Natural Language Understanding System

Authors: DU Jia-li, YU Ping-fang

PAGE 1 – 6

Paper 2: Extending UML for trajectory data warehouses conceptual modelling

Authors: Wided Ouesatli, Jalel Akaichi

PAGE 7 – 12

Paper 3: Optimal itinerary planning for mobile multiple agents in WSN

Authors: Mostefa BENDJIMA, Mohamed FEHAM

PAGE 13 – 19

Paper 4: Genetic procedure for the Single Straddle Carrier Routing Problem

Authors: Khaled MILI, Faissal MILI

PAGE 20–25

Paper 5: The Virtual Enterprise Network based on IPsec VPN Solutions and Management

Authors: Sebastian Marius Rosu, Marius Marian Popescu, George Dragoi, Ioana Raluca Guica

PAGE 26– 34

Paper 6: Analyzing the Efficiency of Text-to-Image Encryption Algorithm

Authors: Ahmad Abusukhon, Mohammad Talib, Maher A. Nabulsi

PAGE 35 – 38

Paper 7: DNA Sequence Representation and Comparison Based on Quaternion Number System

Authors: Hsuan-T. Chang, Chung J. Kuo, Neng-Wen Lo, Wei-Z.Lv

PAGE 39–46

Paper 8: Evaluation of Regressive Analysis Based Sea Surface Temperature Estimation Accuracy with NCEP/GDAS Data

Authors: Kohei Arai

PAGE 47 – 52

Paper 9: The Modelling Process of a Paper Folding Problem in GeoGebra 3D

Authors: Muharrem Aktumen, Bekir Kursat Doruk, Tolga Kabaca

PAGE 53– 57

Paper 10: Sensitivity Analysis of Fourier Transformation Spectrometer: FTS against Observation Noise on Retrievals of Carbon Dioxide and Methane

Authors: Kohei Arai, Hiroshi Okumura, Takuya Fukamachi, Shuji Kawakami, Hirofumi Ohyamai

PAGE 58– 64

Paper 11: Sensitivity Analysis for Water Vapour Profile Estimation with Infrared: IR Sounder Data Based on Inversion

Authors: Kohei Arai

PAGE 65– 70

Paper 12: Wavelet Based Change Detection for Four Dimensional Assimilation Data in Space and Time Domains

Authors: Kohei Arai

PAGE 71– 75

Paper 13: Method for Image Source Separation by Means of Independent Component Analysis: ICA, Maximum Entropy Method: MEM, and Wavelet Based Method: WBM

Authors: Kohei Arai

PAGE 76 –81

Paper 14: Multifinger Feature Level Fusion Based Fingerprint Identification

Authors: Praveen N, Tessamma Thomas

PAGE 82 – 88

Paper 15: Gender Effect Canonicalization for Bangla ASR

Authors: B.K.M. Mizanur Rahman, Bulbul Ahamed, Md. Asfak-Ur-Rahman, Khaled Mahmud, Mohammad Nurul Huda

PAGE 89–94

Paper 16: Three Layer Hierarchical Model for Chord

Authors: Waqas A. Imtiaz, Shimul Shil, A.K.M Mahfuzur Rahman

PAGE 95 – 99

Paper 17: Automatic Facial Expression Recognition Based on Hybrid Approach

Authors: Ali K. K. Bermani, Atef Z. Ghalwash, Aliaa A. A. Youssif

PAGE 100 – 105

Paper 18: Design of A high performance low-power consumption discrete time Second order Sigma-Delta modulator used for Analog to Digital Converter

Authors: Radwene LAJIMI, Mohamed MASMOUDI

PAGE 106 – 112

Paper 19: Short Answer Grading Using String Similarity and Corpus-Based Similarity

Authors: Wael H. Gomaa, Aly A. Fahmy

PAGE 113 – 119

Paper 20: Hybrid intelligent system for Sale Forecasting using Delphi and adaptive Fuzzy Back-Propagation Neural Networks

Authors: Attarius Hicham, Bouhorma Mohammed, Sofi Anas

PAGE 120 – 128

Paper 21: Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches

Authors: Md. Mijanur Rahman, Md. Al-Amin Bhuiyan

PAGE 129 – 137

Paper 22: A Randomized Fully Polynomial-time Approximation Scheme for Weighted Perfect Matching in the Plane

Authors: Yasser M. Abd El-Latif, Salwa M. Ali, Hanaa A.E. Essa, Soheir M. Khamis

PAGE 138 – 143

Paper 23: Cost Analysis of Algorithm Based Billboard Manger Based Handover Method in LEO satellite Networks

Authors: Suman Kumar Sikdar, Soumaya Das, Debabrata Sarddar

PAGE 144 – 150

Paper 24: An agent based approach for simulating complex systems with spatial dynamics application in the land use planning

Authors: Fatimazahra BARRAMOU, Malika ADDOU

PAGE 151 –157

Paper 25: Improving Web Page Prediction Using Default Rule Selection

Authors: Thanakorn Pamutha, Chom Kimpan Siriporn Chimplee, Parinya Sanguansat

PAGE 158– 163

Paper 26: Multilayer Neural Networks and Nearest Neighbour Classifier Performances for Image Annotation

Authors: Mustapha OUJAOURA, Brahim MINAOUI, Mohammed FAKIR

PAGE 164 – 170

Paper 27: Minimization of Call Blocking Probability using Mobile Node velocity

Authors: Suman Kumar Sikdar, Uttam Kumar Kundu, Debabrata Sarddar

PAGE 171 – 179

Paper 28: A Pheromone Based Model for Ant Based Clustering

Authors: Saroj Bala, S. I. Ahson, R. P. Agarwal

PAGE 180– 183

Paper 29: Modern and Digitalized USB Device with Extendable Memory Capacity

Authors: J.Nandini Meeraa, S.Devi Abirami, N.Indhuja, R.Aravind, C.Chithiraikkayalvizhi, K.Rathina Kumar

PAGE 184 – 188

Paper 30: Enhanced Modified Security Framework for Nigeria Cashless E-payment System

Authors: Fidelis C. Obodoeze , Francis A. Okoye

PAGE 189 – 196

Paper 31: Cashless Society: An Implication To Economic Growth & Development In Nigeria

Authors: N. A. YEKINI,I. K. OYEYINKA, O.O. AYANNUGA, O. N. LAWAL, and A. K. AKINWOLE

PAGE 197 – 203

Paper 32: The Analysis of the Components of Project-Based Learning on Social Network

Authors: Kuntida Thamwipat, Napassawan Yookong

PAGE 204 – 209

Paper 33: A Database Creation for Storing Electronic Documents and Research of the Staff

Authors: Pornpapatsorn Princhankol, Siriwan Phacharakreangchai

PAGE 210– 214

Predicting Garden Path Sentences Based on Natural Language Understanding System

DU Jia-li

School of Foreign Languages/School of Literature
Ludong University/Communication University of China
Yantai/ Beijing, China

YU Ping-fang

School of Liberal Arts/Institute of Linguistics
Ludong University/Chinese Academy of Social Sciences
Yantai/ Beijing, China

Abstract— Natural language understanding (NLU) focusing on machine reading comprehension is a branch of natural language processing (NLP). The domain of the developing NLU system covers from sentence decoding to text understanding and the automatic decoding of GP sentence belongs to the domain of NLU system. GP sentence is a special linguistic phenomenon in which processing breakdown and backtracking are two key features. If the syntax-based system can present the special features of GP sentence and decode GP sentence completely and perfectly, NLU system can improve the effectiveness and develop the understanding skill greatly. On the one hand, by means of showing Octav Popescu's model of NLU system, we argue that the emphasis on the integration of syntactic, semantic and cognitive backgrounds in system is necessary. On the other hand, we focus on the programming skill of IF-THEN-ELSE statement used in N-S flowchart and highlight the function of context free grammar (CFG) created to decode GP sentence. On the basis of example-based analysis, we reach the conclusion that syntax-based machine comprehension is technically feasible and semantically acceptable, and that N-S flowchart and CFG can help NLU system present the decoding procedure of GP sentence successfully. In short, syntax-based NLU system can bring a deeper understanding of GP sentence and thus paves the way for further development of syntax-based natural language processing and artificial intelligence.

Keywords- Natural language understanding; N-S flowchart; computational linguistics; context free grammar; garden path sentences.

I. INTRODUCTION

Natural language understanding (NLU), speech segmentation, text segmentation, part-of-speech tagging, word sense disambiguation, syntactic ambiguity, etc. come under the umbrella term "natural language processing (NLP)".[1] The development of NLU is briefly traced from the early years of machine translation to today's question answering and translation systems. [2-3]NLU today deals with machine reading comprehension in artificial intelligence (AI) [4]and is applied to a diverse set of computer applications.[5-6] Its subject ranges from simple tasks such as short commands issued to robots, to complex endeavors such as the full comprehension of articles or essays. A machine created to understand natural language has been one of the dreams of AI ever since computers were invented, and this encourages the systematic and rapid development of NLU. The efficient understanding of natural language requires that computer program be able to resolve ambiguities at the syntactic level

and recover that part of the meaning of its individual words taken in isolation.[7-8] The satisfaction of this requirement involves complex inference from a large database of world-knowledge, and this makes the designer of computer programs for NLU face the serious difficulty of algorithm processing.[9] The machine comprehension is embedded in the more general frame of interpersonal communication and is applied to the person-machine interaction task.[10] The further integration has proved appropriate for the design of effective and robust natural language interfaces to AI systems. Syntactic garden path phenomenon is a major source of uncertainty in NLU. Syntax-supported systems attempt to help machine to get a deeper understanding of garden path sentence and the related algorithms deserve special attention in the future of NLU developing. [11] Many hybrid researches contribute to the improvement of NLU systems. [12] For example, the fact is established that abduction rather than deduction is generally viewed as a promising way to apply reasoning in NLU. [13]The development of NLU can focus on the design of a stochastic model topology that is optimally adapted in quality and complexity to the task model and the available training data.[14] A comparative information-theoretic study is carried out to show positional letter analyses, n-gram analyses, word analyses, empirical semantic correlations between the Greek and English n-grams, and entropy calculations are useful to text processing and compression, speech synthesis and recognition as well as error detection and correction.[15]

Garden path (GP) phenomenon is a special linguistic phenomenon which comprises processing breakdown and backtracking. GP sentence (e.g. "The old dog the footsteps of the young") is an originally correct sentence which makes readers' grammatical misinterpretation linger until re-decoding has occurred. An incorrect choice in GP sentence usually is readers' most likely interpretation, leading readers initially into an improper parse which, however, finally proves to be a dead end. Thus the processing breaks down and backtrack to the original status to search the given information again for alternative route to the successful decoding. The automatic decoding of GP sentence is a challenge for NLU systems for machine has to have access to grammatical, semantic and cognitive knowledge in order to understand natural language smoothly as human brains do.

In this paper, a syntax-based and algorithm-originated approach to understanding GP sentence in natural language domain is presented. The discussion consists of four sections. Firstly, a logic-based model of NLU system is shown to

provide an overview of NLU. Secondly, N-S flowchart is used to present the special feature of processing breakdown of GP sentence. Thirdly, context-free grammar (CFG) is introduced to analyze the backtracking of GP sentence in detail and the machine's automatic decoding procedure of GP sentence is analyzed. The last section brings the conclusion.

II. A MODEL OF NATURAL LANGUAGE UNDERSTANDING SYSTEM

In the domain of Intelligent Tutoring Systems, the high-precision NLU system is helpful to accurately determine the semantic content of learners' explanations. For example, the NLU system developed in the context of the Geometry Cognitive Tutor combines unification-based syntactic processing with Description Logic based semantics to achieve the necessary accuracy level. [16] The syntactic and semantic processing of natural language is useful for specific decoding problems, like metonymy resolution and reference resolution. On the basis of linguistic theories and computational technologies, NLU system architecture building the syntactic structure and offering the semantic interpretation of learner' explanations is structured by Octav Popescu in 2005. (Fig. 1)

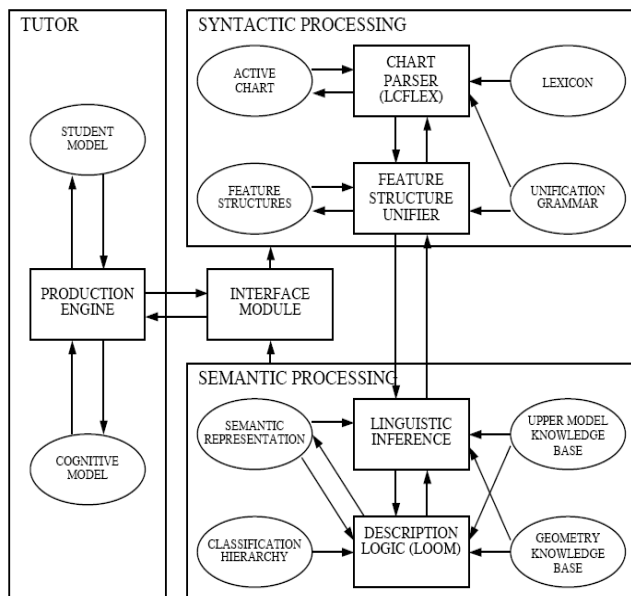


Figure. 1. Octav Popescu's NLU system architecture

The research is related to practical application of NLU system. Popescu's architecture comprises two sections: the syntactic and the semantic processing subsystems, which are interactive during the NLU processing. The syntactic processing subsystem uses an active chart parser while the semantic one bases its action on a Description Logic system. The key part is interface module connecting the NLU subsystem to the tutor itself, functioning asynchronously to the NLU system, taking the input sentence from the tutor, and passing the checked words to the chart parser. System functions in real time until the parser finishes and passes the resulting classifications back to the tutor. The chart parser, adopting linguistic knowledge about the target natural language from the unification grammar and the lexicon, is the center of NLU system. Words of a sentence are taken one by

one by the parser according to grammar rules, and feature structures, which can store lexical, syntactic, and semantic properties of corresponding words and phrases, are built simultaneously.

Popescu's system highlights the integration of syntactic and semantic information during the machine decoding, showing that hybrid knowledge is necessary for machine to understand natural language. GP sentence is a more complex linguistic phenomenon than the common sentence since GP sentence possesses special features of processing breakdown and backtracking. Therefore, the emphasis of integration of linguistics and computational science is helpful for system to automatically decode GP sentence. The computational skill (e.g. N-S flowchart) can present in detail the syntactic processing procedure, making learners have access to the machine understanding of GP sentence and thus enhancing the effectiveness of system.

III. N-S FLOWCHART BASED NATURAL LANGUAGE UNDERSTANDING OF GP SENTENCE

GP sentence seems to be a grammatically correct sentence at the original processing stage and is usually used in syntax, linguistics, psycholinguistics, and computational linguistics. According to syntactic theory, while a person reads a GP sentence, he builds up an original meaning structure and processes natural language one word at a time. With the advancement of processing, the person finds that he has been constructing an incorrect structure and that the next word or phrase cannot be incorporated into the structure originally created thus far. The "garden path" is a reference to the saying "to be led down the garden path", meaning "to be misled". Examples are as follows: "The horse raced past the barn fell", "The man who hunts ducks out on weekends", "The cotton clothing is made of grows in Mississippi", "The prime number few", "Fat people eat accumulates", "The old man the boat", "The tycoon sold the offshore oil tracts for a lot of money wanted to kill JR", etc (Some examples cited are from <http://home.tiac.net/~cri/1998/garpath1.html>). The verbs (e.g., "raced", "ducks", "grows", "number", "accumulates") in the above examples are not the words which a reader is assuming to be the verbs. Some words that look as though they were modifying nouns are actually being used as nouns and some words which look as though they were nouns are verbs. The revised non-GP sentences are as follows: "The horse that was raced past the barn fell down", "The man, who hunts, ducks out on weekends.", "The cotton, of which clothing is made, grows in Mississippi", "The prime (group) number few", "Fat that people eat accumulates", "The old (people) man the boat", "The tycoon, who was sold the offshore oil tracts for a lot of money, wanted to kill JR", etc. Please see the explanation in Fig. 2.

Fig. 2 shows us that GP sentence which has special features of processing breakdown and backtracking is exclusive of multi-meanings, namely, only one meaning involved in the final result of decoding even though the deviation of the original meaning and the final meaning is obvious. The special features of GP sentence can be presented with the help of computational linguistics skills, e.g. N-S flowchart which was raised by Nassi and Shneiderman [17] to

analyze structured programming [18-19]. This kind of analysis is helpful for the effectiveness of discussion boards. [20]

Sentence	Initial likely partial parse	Final parse	Alternative form of original sentence
The horse raced past the barn fell.	The horse was racing past the barn....	The horse that was raced past the barn fell down.	The horse was raced past the barn and fell down.
The man who hunts ducks out on weekends.	The man who hunts ducks....	The man, who hunts, ducks out on weekends.	The man hunts and ducks out on weekends.
The cotton clothing is made of grows in Mississippi.	The clothing, which is made of cotton, is made of....	The cotton, of which clothing is made, grows in Mississippi.	The cotton that clothing is made of grows in Mississippi.
The prime number few.	The prime number....	The prime (group) number few.	The major group count few.
Fat people eat accumulates.	Fat people eat....	Fat that people eat accumulates.	An oily substance (fat) that people eat accumulates
The old man the boat.	The old man....	The old (people) man the boat.	The old people sail the boat.
The tycoon sold the offshore oil tracts for a lot of money wanted to kill JR.	The tycoon sold the offshore oil tracts for a lot of money....	The tycoon, who was sold the offshore oil tracts for a lot of money, wanted to kill JR.	The tycoon was sold the offshore oil tracts for a lot of money and wanted to kill JR.

Figure. 2. Explanation for garden path sentence

Flowchart’s IF-THEN-ELSE statement in computational processing can be used to show the procedure of decoding GP sentence now that this technique adopts a dichotomy between Yes and No. For example, Fig.3 consists of four pairs of dichotomies, namely, A and not A, B and not B, C and not C, D and not D. Nassi and Shneiderman introduce seven possible processing results, i.e. (1) not A; (2) not A and not B; (3) not A and B; (4) A and not C; (5) A and not C and not D; (6) A and not C and D; (7) A and C.

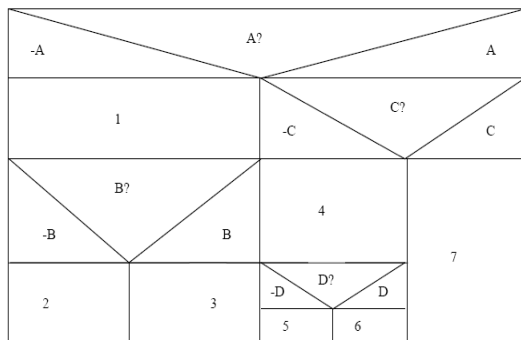


Figure. 3 Nassi and Shneiderman’s flowchart of IF-THEN-ELSE statement

The dichotomy of IF-THEN-ELSE statement can be applied to analysis of GP sentence which possesses processing breakdown and backtracking.

Example 1. The new record the song.

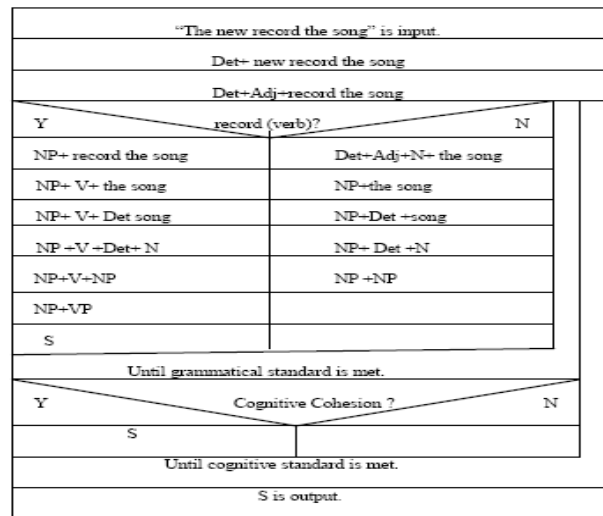


Figure. 4 The example of two pairs of dichotomies of N-S flowchart

There are two pairs of dichotomies involved in Fig. 4. The first pair is used to distinguish the verb definition and the noun definition of “record”. If the verb definition is chosen, the machine decoding will run along the left column which is the grammatically correct presentation.

Otherwise, the noun definition of “record” is the alternative choice; then the right column is the preferable choice. According to the cognitive rules, “Det+Adj+Noun” is the prototype model in which “Adj” is used to modify the “Noun”. Therefore, the right column is the original decoding procedure until system fails to parse more. The decoding has to return to the triangle to choose the verb definition of “record” till every string involved is processed completely.

The second pair of dichotomy appears after the grammatical requirements are met successfully. That is to say, system usually abides by the grammatical rules firstly for the syntactic cohesion is the basic condition for decoding GP sentence.

After satisfying the grammatical requirements, system then considers cognitive cohesion. This is the reason why cognitive rules are put after the grammatical decoding. Example 1 is accepted by system according to cognitive rules and therefore the result is output finally. If the sentence “The dead record the song” which has the similar decoding procedure like Example 1 is submitted to system, the cognitive disagreement makes system reject the decoding and no final result will be output.

Example 2. The old make the young man the boat.

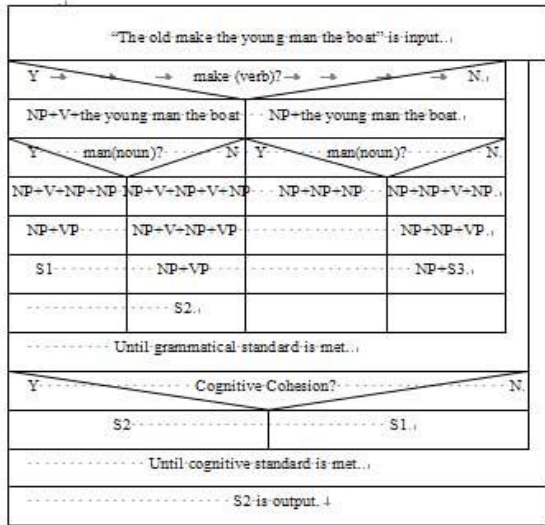


Figure.5 The example of four pairs of dichotomies of N-S flowchart

Fig. 5 consists of four pairs of dichotomies of N-S flowchart. The first pair is used to differentiate the verb definition from the noun definition of “make”. If the verb definition is chosen, the left column is system’s decoding path. Otherwise, the noun definition in the right column is the choice. The second central triangle in left column is applied to distinguish the verb definition from the noun definition of “man”. The result is that the noun one is in left part and the verb one is in right part.

The third pair in right column is similar decoding procedure as the second pair. The difference between the second and the third lies in “make” in the former is a verb while “make” in the latter is a noun. The last pair of dichotomy is used to test the output from the cognitive level. If the result is for the cognitive rules, the left column is activated and it will be output successfully. Otherwise, the right column starts and the failed system has to return to the first pair of dichotomy to parse again.

For example, “S1” is a grammatical correct output of Example 2, and it means that “The old people make the young man be the boat” which is against the cognitive rules. Therefore, “S1” is rejected by system and has to return to the first central triangle until the correct “S2” is chosen. “S2” means “The old people make the young people sail the boat”.

From the discussion above, we can find that N-S flowchart is useful to analyze GP sentence since this technique has the obvious function by which the special features of processing breakdown and backtracking can be presented in detail.

Besides N-S flowchart, N. Chomsky’s context-free grammar (CFG) is often applied to natural language understanding and the formalized processing is helpful in decoding GP sentence.

IV. CFG-BASED NATURAL LANGUAGE UNDERSTANDING OF GP SENTENCE

A context-free grammar (CFG) is also called a phrase structure grammar in formal language theory. According to the

grammar, it can naturally generate a formal language and clauses can be embedded inside clauses arbitrarily deeply. In terms of production rules, every production of a CFG is of the form: “V→w”. Generally speaking, “V” is a single nonterminal symbol, and “w” is a string of terminals and/or nonterminals. CFG used to analyze the syntax of natural languages plays an important role in the description and design of programming languages in NLU system.

Noam Chomsky has argued that natural languages are based on CFG and the processing procedure can be presented by formal languages. In Chomsky’s opinion, if “A→BC” or “A→a” can be accepted by system in which “A (B.C)” represents the nonterminal symbol and “a” is regarded as terminal symbol, the CFG (usually called Chomsky Normal Form) is a dichotomy possessing the binary tree form. Thus the programming languages are designed to decode GP sentence.

According to formal languages, quaternion parameters are involved in the programming, namely, “Vn” (the collection of nonterminal symbols), “Vt” (the collection of terminal symbols), “S” (the original start), and “P” (rewriting programs). Thus, the natural language can be understood by a machine if the programming language is created. The analysis of GP sentence of Example 2 is as follows in Fig.6.

- G={Vn, Vt, S, P}
- Vn={S, NP, VP, Det, Adj, V, N}
- Vt={the, old, make, young, man, boat}
- S=S
- P:

 - S → NP VP (a)
 - NP → Det Adj (b)
 - NP → Det N (c)
 - NP → Det Adj N (d)
 - VP → V NP NP (e)
 - VP → V NP VP (f)
 - VP → V NP (g)
 - Det → {the} (h)
 - Adj → {old, young} (i)
 - N → {make, man, boat} (j)
 - V → {make, man} (k)

Processing Procedure (Part I: NP+NP+NP)

1. The old make the young man the boat
2. Det old make the young man the boat (h)
3. Det Adj make the young man the boat (i)
4. Det Adj N the young man the boat (j)
5. NP the young man the boat (d)
6. NP Det young man the boat (b)
7. NP Det Adj man the boat (i)
8. NP Det Adj N the boat (j)
9. NP NP the boat (d)
10. NP NP Det boat (h)
11. NP NP Det N (j)
12. NP NP NP (c)

Processing Procedure (Part II: NP+S3)

- 13(7). NP Det Adj man the boat
14. NP NP man the boat (b)
15. NP NP V the boat (k)
16. NP NP V Det boat (h)
17. NP NP V Det N (j)
18. NP NP V NP (c)
19. NP NP VP (g)
20. NP S3 (a)

Processing Procedure (Part III: S1)	
21(3). Det Adj make the young man the boat	
22. NP make the young man the boat	(b)
23. NP V the young man the boat	(k)
24. NP V Det young man the boat	(h)
25. NP V Det Adj man the boat	(i)
26. NP V Det Adj N the boat	(j)
27. NP V NP the boat	(d)
28. NP V NP Det boat	(h)
29. NP V NP Det N	(j)
30. NP V NP NP	(c)
31. NP VP	(e)
32. S1	(a)
Processing Procedure (Part IV: S2)	
32(25). NP V Det Adj man the boat	
33. NP V NP man the boat	(b)
34. NP V NP V the boat	(k)
35. NP V NP V Det boat	(h)
36. NP V NP V Det N	(j)
37. NP V NP V NP	(c)
38. NP V NP VP	(e)
39. NP VP	(f)
40. S2	(a)

Figure 6 CFG-based understanding of Example 2

The whole processing procedure of Example 2 is shown above and comprises 40 steps for a machine to understand the GP sentence. The flowchart of the decoding consists of four parts. The first part is the production of “NP+NP+NP”; the second, “NP+S3”; the third, “S1”; the fourth, “S2”.

In the first part, not all strings are decoded completely according to the rewriting rules and correspondingly system fails to go ahead. The decoding returns to the central triangle where “man” is considered a verb or a noun. The verb definition of “man” is chosen since the noun definition proves to be a dead end. Thus, system comes to the second part of processing procedure in which “NP+S3” is the conclusion. The decoding numbers of first part is “1-2-3-4-5-6-7-8-9-10-11-12-12-11-10-9-8-7”.

The second part is the result of backtracking of “NP+NP+NP”. “Man” is alternatively regarded as a verb and system reads the sentence of “The old make the young man the boat” into “NP+S3” in which “make” is considered a noun. Since no more rewriting rules support the systematic decoding of “NP+S3”, the syntactic output is against the grammatical requirements and as a result, system rejects the final output in which both the noun definition and the verb definition of “man” are chosen as the syntactic choices when “make” is used as a noun. Therefore, system has to backtrack to another central triangle where “make” is considered to be a verb, which brings the appearance of the third part. The decoding numbers of second part is “13(7)-14-15-16-17-18-19-20-20-19-18-17-16-15-14-13-12-11-10-9-8-7-6-5-4-3”.

The third part is the result of backtracking of “NP+NP+NP” and “NP+S3” where noun definition of “make” is preferable choice now that, according to cognitive and psychological prototype theory, “Det+Adj+Noun” is the default setting. Since the noun definition of “make” results in a dead end, system becomes to consider “make” a verb. Thus the left central triangle is activated. If “man” is chosen as a noun, S1 in which Example 2 means “The old people make the young man (be) the boat” is the temporary decoding result. There is no cognitive cohesion involved in S1 even though it

is grammatical correct. System rejects S1 as a successful output (see Fig. 5) and backtracks to the central triangle where “man” can be chosen as a verb, which results in the decoding of S2. The decoding numbers of third part is “21(3)-22-23-24-25-26-27-28-29-30-31-32-32-31-30-29-28-27-26-25”.

The last part is the result of backtracking of “NP+NP+NP”, “NP+S3” and “S1”. The only choice for system is that both “make” and “man” should be verbs. The meaning of Example 2 is “The old people make the young people sail the boat”. This explanation satisfies the grammatical, cognitive and semantic requirements and is accepted by system perfectly. The automatic understanding of GP sentence in NLU system is completely concluded. The decoding numbers of last part is “32(25)-33-34-35-36-37-38-39-40”.

CFG-based natural language understanding is helpful to analyze the special features of GP sentence. Both processing breakdown and backtracking can be presented on the basis of CFG grammar. For example, in the decoding of Example 2, all the figures in parentheses refer to the stage in which processing breaks down and the converse decoding numbers represent the stage in which backtracking occurs. The whole decoding numbers of NLU system is as follows: “1-2-3-4-5-6-7-8-9-10-11-12-12-11-10-9-8-7-13(7)-14-15-16-17-18-19-20-20-19-18-17-16-15-14-13-12-11-10-9-8-7-6-5-4-3-21(3)-22-23-24-25-26-27-28-29-30-31-32-32-31-30-29-28-27-26-25-32(25)-33-34-35-36-37-38-39-40”.

V. CONCLUSION

Natural language understanding (NLU) dealing with machine reading comprehension is the umbrella term of “natural language processing (NLP)”. The advancement of machine technologies and computational skills develops NLU system. Its subject ranges from sentence decoding to text understanding. The decoding of GP sentence, a special linguistic phenomenon which possesses processing breakdown and backtracking, belongs to the domain of NLU system. GP sentence is a grammatical correct sentence in which original misinterpretation lingers until re-decoding has occurred. An incorrect choice in GP sentence usually is readers' most likely interpretation and lures readers into a default parse which, however, finally proves to be a dead end. If the system can present the special features of GP sentence, namely, processing breakdown and backtracking, the automatic decoding can be successful and the effectiveness of system can be improved. After showing Octav Popescu's model of NLU system, we emphasize the importance of integration of syntactic, semantic and cognitive backgrounds in system, focus on the programming skill of IF-THEN-ELSE statement used in N-S flowchart, and highlight the function of context free grammar (CFG) created to decode GP sentence. On the basis of GP sentences-supported analyses, we come to the conclusions that the machine comprehension can be embedded in the general frame of communication via programming languages, that interaction between the person and machine can improve NLU system, that programming technology (e.g. N-S flowchart) can help NLU system present the decoding procedure of GP sentence, and that syntax-supported linguistic skill (e.g. CFG) can bring a deeper understanding of GP sentence.

ACKNOWLEDGMENT

This research is supported in part by grants of YB115-29 and YB115-41 from “the Eleventh Five-year” research projects of Chinese Language Application, and grant 11YJA740111 from the Ministry of Education and Science Planning Project. This research is also sponsored by 2012 Yantai Social Science Planning Project “Towards the Linguistic Function of the Cultural Development in Yantai: A Perspective of Computational Linguistics”.

REFERENCES

- [1] W. L. Wu, R. Z. Lu, J. Y. Duan, H. Liu, F. Gao, and Y. Q. Chen, “Spoken language understanding using weakly supervised learning,” *Computer Speech & Language*, vol. 24, April 2010, pp. 358-382.
- [2] Y. Wilks, “A position note on natural language understanding and artificial intelligence,” *Cognition*, vol. 10, July-August 1981, pp. 337-340.
- [3] R. C. Schank, “Conceptual dependency: A theory of natural language understanding,” *Cognitive Psychology*, vol. 3, October 1972, pp. 552-631.
- [4] C. Mellish, and J. Z. Pan, “Natural language directed inference from ontologies,” *Artificial Intelligence*, vol. 172, June 2008, pp. 1285-1315.
- [5] M. Jeong, and G. G. Lee, “Practical use of non-local features for statistical spoken language understanding,” *Computer Speech & Language*, vol. 22, April 2008, pp. 148-170.
- [6] M. Jeong and G. G. Lee, “Machine learning approaches to spoken language understanding for ambient intelligence,” *Human-Centric Interfaces for Ambient Intelligence*, 2010, pp. 185-224.
- [7] K. Hird and K. Kirsner, “Objective measurement of fluency in natural language production: A dynamic systems approach,” *Journal of Neurolinguistics*, March 2010.
- [8] A. Lockman, and D. Klappholz, “The control of inferencing in natural language understanding,” *Computers & Mathematics with Applications*, vol. 9, 1983, pp. 59-70.
- [9] R. López-Cózar, Z. Callejas, and D. Griol, “Using knowledge of misunderstandings to increase the robustness of spoken dialogue systems,” *Knowledge-Based Systems*, vol. 23, July 2010, pp. 471-485.
- [10] C. Madden, M. Hoen, and P. F. Dominey, “A cognitive neuroscience perspective on embodied language for human-robot cooperation,” *Brain and Language*, vol. 112, March 2010, pp.180-188.
- [11] K. J. Lee, Y. S. Choi, and J. E. Kim, “Building an automated English sentence evaluation system for students learning English as a second language,” *Computer Speech & Language*, May 2010.
- [12] D. Kayser, and F. Nouioua, “From the textual description of an accident to its causes,” *Artificial Intelligence*, vol. 173, August 2009, pp. 1154-1193.
- [13] J. Bos, “Applying automated deduction to natural language understanding,” *Journal of Applied Logic*, vol. 7, March 2009, pp. 100-112.
- [14] W. Minker, “Design considerations for knowledge source representations of a stochastically-based natural language understanding component,” *Speech Communication*, vol. 28, June 1999, pp. 141-154.
- [15] E. J. Yannakoudakis, I. Tsomokos, and P.J. Hutton, “N-grams and their implication to natural language understanding,” *Pattern Recognition*, vol. 23, 1990, pp. 509-528.
- [16] O. Popescu, *Logic-Based Natural Language Understanding in Intelligent Tutoring Systems* (PhD Thesis), Carnegie Mellon University, 2005.
- [17] I. Nassi and B. Shneiderman, “Flowchart techniques for structured programming,” *ACM SIGPLAN Notices*, vol. 8, 1973, pp. 12-26.
- [18] J. L. Du and P F Yu, “A computational linguistic approach to natural language processing with applications to garden path sentences analysis,” *International Journal of Advanced Computer Science and Applications*. Vol. 3, September 2012, pp. 61-75.
- [19] J. L. Du, P. F. Yu, “Syntax-directed machine translation of natural language: Effect of garden path phenomenon on sentence structure,” *International Conference on Intelligent Systems Design and Engineering Applications*, 2010, pp. 535–539.
- [20] A. A. Badawy, “Students’ perceptions of the effectiveness of discussion boards,” *International Journal of Advanced Computer Science and Applications*. Vol. 3, September 2012, pp. 136-144.

Extending UML for trajectory data warehouses conceptual modelling

Wided Oueslati^{#1}, Jalel Akaichi^{*2}

Computer Science Department, Higher Institute of Management
Tunisia

ABSTRACT— The new positioning and information capture technologies are able to treat data related to moving objects taking place in targeted phenomena. This gave birth to a new data source type called trajectory data (TD) which handle information related to moving objects. Trajectory Data must be integrated in a new data warehouse type called trajectory data warehouse (TDW) that is essential to model and to implement in order to analyze and understand the nature and the behavior of movements of objects in various contexts. However, classical conceptual modeling does not incorporate the specificity of trajectory data due to the complexity of their components that are spatial, temporal and thematic (semantic). For this reason, we focus in this paper on presenting the conceptual modeling of the trajectory data warehouse by defining a new profile using the StarUML extensibility mechanism.

Keywords—Trajectory data; trajectory data warehouse; UML extension; trajectory UML profile.

I. INTRODUCTION

The success of the warehousing process rests on a good conceptual modeling schema. In fact, conceptual modeling offers a higher level of abstraction while describing the data warehousing project since it stays valid in case of technological evolution. Besides, it allows determining analysis possibilities for the warehouse. However, no contribution is at the present time standard in term of trajectory data semantic models. This finding leads us to propose a new UML profile with user oriented graphical support to represent trajectory data and trajectory data warehouse conceptual modeling with structural model (class diagram) and dynamic model (sequence diagram).

This paper is organized as follows. In section 2, we present an overview of research works related to conceptual approaches and extensibility of UML for applications' needs. In section 3, we present the methodology that we adopted to extend the StarUML profile. In section 4, we present the Trajectory UML profile. In section 5, we present the trajectory UML profile realization. In section 6, we summarize the work and we propose some perspectives that can be done in the future.

II. RELATED WORKS

In this section, we present different approaches related to the conceptual modeling methodology, then we present research works that extended UML to adopt it to their conceptual modeling needs. In the literature, we can find three categories of conceptual approaches; the top down approach, the bottom up approach and the middle out approach. The

difference between those latter is situated in the starting point. In fact, each approach has its own starting point such as users' needs, data marts or both users' needs and data marts. Concerning the top down approach, this latter has to answer users' requirements without any exception. It is very expensive in term of time since it requires the whole conceptual modeling of the DW as well as its realization and it is difficult because it requires the knowledge in advance of dimensions and facts [1]. In this category, authors of [2] present a Multidimensional Aggregation Cube (MAC) method. This latter insures the construction of a multidimensional schema from the definition of decision makers' needs but the defined schema is partial because it describes only the hierarchies of dimensions. The goal of MAC is to supply an intuitive methodology of data modeling used in the multidimensional analysis. It models real world scenarios using concepts which are very similar to OLAP.

In MAC, data are described as dimensional levels, drilling relationships, dimensions, cubes and attributes. Dimension levels are a set of dimension members. Those latter are the most detailed modeling concepts and they present real world instances' properties. Drilling relationships are used to present how one level element can be decomposed of other levels' elements. The dimension paths present a set of drilling relationships which are used to model a significant sequence of drill down operations. Dimensions are used to define a significant group of dimension paths. This grouping is essential to model semantic relationships. Cubes are the only concept which associates properties' values with real measures' values. They insist on the complex hierarchy structure defined by dimensions. The top down approach can be used in the Goal-driven methodology [3]. In fact, this latter focuses on the company's strategy by occurring the executives of the company. For the bottom up approach, this latter consists on creating the schema step by step (data marts) until the obtaining of a real DW [1].

It is simple to be realized but it requires an important work in the data integration phase. Besides, there is always the risk of redundancy due to the fact that each table is created independently. Authors in [4] present a dimensional fact model. This latter relays on the construction of data marts firstly. This can insure the success in case of complex projects but it neglects the role of decision makers. Authors in [5] adopt the bottom up approach. In fact, they present a dimensional model development method from traditional Entity-Relationship models to insure the modelling of DWs and Data Marts. This method is based on three steps: the first step includes the classification of data models' entities into a

set of categories. This leads to the production of a dimensional model from an Entity-Relationship model. We find the transactional entities that insure the storage of details concerning particular events in the company. We find also the component entities that are directly connected to transactional entities through the 1..* relationship. Those entities allow defining details of each transaction. Classification entities are connected to component entities through the 1..* relationship. Classification entities present the existing hierarchies in the model.

The second step consists on identifying hierarchies that exist in the model. In fact, the hierarchy is an important concept in the dimensional modelling level. The third step consists on grouping hierarchies and aggregations together to form a dimensional model. At this level, we find two operators that are used to product models of dimensions. In fact, the first operator can transform the high level entities to low level entities. This can be done until the arrival at the bottom of the architecture. The aim is to have an only table at the end. For the second operator, it is applied on transactional data to create a new entity which contains summarized data. This approach is used as a base for the data driven and user driven methodologies. In fact, as presented in [3], the data driven (supply driven) methodology starts by analyzing operational data sources to identify existent data. Users' intervention is limited to the choice of necessary data for the decision making process. This methodology is adopted when data sources are valid.

For the user driven methodology, it starts by collecting users' needs. Those needs will be integrated in order to obtain one multidimensional schema. This approach is appreciated by users but it presents a big challenge. In fact, managers of projects must be able to take into account the different points of views. For the middle out approach, it is an hybrid method since it benefits from the two approaches cited above. Authors in [6] present an example of hybrid modelling method that is based on the top down and the bottom up approaches. The bottom up approach is based on three steps: the collection of needs, the specification and the formalization of those needs in the form of multidimensional constellation schema. The top down approach includes the data collection and the construction of a multidimensional schema that allows decision making. The approach is based on the description of decision makers needs. Those two approaches allow having two schemas, then from those latter only one schema will be derived and kept. The middle out approach is composed of four phases; the users' needs analysis, the confrontation/comparison, the resolution of conflicts and the implementation.

Authors in [7] present another method which uses the middle out approach. This latter is based on three steps: the collection of users' requirements by the top down approach, the recovery of star schema by the bottom up approach and finally the integration phase. This latter connects the obtained star schema from the first step to the obtained star schema from the second step. The integration is realized thanks to a set of matrix. Users' requirements are collected by the Goal Question Metric (GQM) paradigm. This latter allow attributing metrics to identified goals. This facilitates the

filtering and the deletion of not useful goals. Authors of [7] consider that the modeling of warehouses is a process based on goals, and then users' goals related to DW development will be present explicitly. Goals will be analyzed in order to reduce their number (authors take into account the similarity of goals). For the choice of star schema, authors use the Entity-Relationship model. This latter is exhaustively analyzed to find entities that will be transformed to facts and dimensions. The transformation process of Entity-Relationship model to a star schema is based on three steps. The first step is the construction of a connected graph that serves to synthesized data. The second step is to extract a snowflake schema from the graph. The third step is the integration phase. In fact, authors exploit the structure of the warehouse of the first phase and the set of possible schemas of the second phase, and then they apply a set of steps such as converting of schema to express them with the same terminology. Within UML-based conceptual models, the most famous approaches are of Trujillo and his team.

In [8], authors proposed UML extensions for object-oriented multidimensional modeling. This extension is performed thanks to stereotype mechanism, tagged values and constraints expressed in OCL-Object Constraint Language, in addition to a set of Well-Formedness rules managing new elements added and determining the semantic of the model. Stereotypes and icons allow an expressive representation of different constituent elements of a multidimensional model namely fact classes, dimension classes, hierarchy levels and attributes. Dimension level classes (stereotyped base classes) should define a directed acyclic graph rooted in dimension class. Concerning relationships, the aggregation links facts to dimensions, and association/generalization links dimension levels (having Base stereotype) between each other's.

In another work of the same team [9], an UML package is proposed to facilitate modeling of large data warehouse systems. In fact, they suggest a set of UML diagrams (package) extended with the aforementioned stereotypes, icons and constraints (OCL) to cope with multidimensional modeling and consequently designers will not be limited only to the class diagram.

Several works [10] [11] [12] [13] [14] [15] proceeded by UML profiles to represent their models. In fact, in [11] the proposed profile is used for the oriented agent modeling. In [12] the authors represented a profile for the mobile systems conception. In [13], authors propose a profile for the modeling of association rules of data mining. In [14] authors propose a profile to model data mining with the temporal series in the data warehouse. In [15], authors extended UML to introduce new stereotypes and icons to handle spatial and temporal properties at the conceptual level. This led to the visual modeling tool so called Perceptory.

III. ADOPTED METHODOLOGY

There are three methods of conceptual modeling of DWs; the first one is the top down approach [16] that is based on the needs of the users, the second is the bottom-up approach [5] that begins with the operational data sources and finally the mixed approach [17] that combines the two previous approaches. We used the top down approach in our modeling

phase because we were interested in user's needs. In term of MDA (Model Driven Architecture) [18] our solution is situated in the CIM (Computation Independent Model) level because the models are not inevitably transformed into code. For the abstraction levels [19] (conceptual, logical and physical) our solution is established to cover the conceptual level. Here is a plan showing the position of our solution:

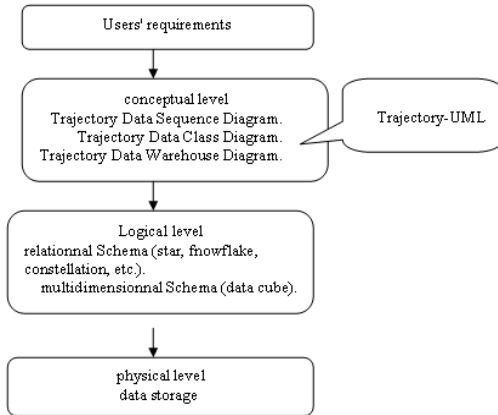


Fig 1. Our solution's position

We also adopted object oriented paradigm because it has several advantages for the multidimensional modeling such as the classification / instantiation, the Generalization / specialization and the Aggregation / decomposition. We chose to adopt the object oriented approach which is based on the UML profiles. We are inspired by the work of [10] to follow the mechanism of profile UML to model the multidimensional part and the work of [20] to widen this profile to the trajectory data and spatiotemporal data.

IV. TRAJECTORY UML PROFILE

With the expansion of technology based on captors (GPS, RFID, etc), it became possible to establish new information systems involving moving objects, to set their trajectories and to pusue their movements. The Trajectory Data Warehouses (TDWs) emerge for needs to study the devices of moving objects in order to develop the decision process. It is necessary to provide a formal representation to TDWs, to understand the concrete world and for a good human comprehension to moving objects phenomena. Modelling TDWs is in its early stage. The classic approaches do not have their good productions for a clear and standard methodology given the great complexity of trajectory data and the difficulty to model varying fields.

An UML profile [21] allows specializing UML in a precise domain, it consists of stereotypes, tagged values and constraints. A stereotype [22] is an element of the model that defines new values, new constraints and a new graphic representation. Its role is to give a semantic representation to an element of the model. A stereotype can be represented as a string character between two quotation marks << >> or with an icon. A marked value specifies a new property attached to an element of the model. It is represented between {} and placed with the name of another element. A constraint can become attached to any element of the model to refine its semantics and prevent an arbitrary use of the various elements.

It can be defined with the natural language and/or with the OCL (object constraint language) [21] which is a declarative language that allows developers to write constraints on the model's objects. Recently, UML profiles have a great progress in the ways for conception of Data Warehouses. We present in this section, a conceptual solution for trajectory data warehouses design. We proceeded by an UML profile in order to add stereotypes, tagged values and constraints. Our Trajectory UML profile contains two diagrams: the first one is Trajectory Data Class Diagram. This latter has for purpose to model the trajectory data of the moving objects. The second diagram is Trajectory Data Warehouse Diagram. This latter represents the TDWs in a multidimensional context.

A. Trajectory Class Diagram

We defined in this diagram stereotypes and icons related to trajectories such as moving object, stop, move, trajectory section, pda, gps and location. This diagram can be used in each case based on trajectories of moving objects.

1) *Classes' Stereotypes:* We defined in this table classes's stereotypes used in the modeling

TABLE I : STEREOTYPES DESCRIPTION AND REPRESENTATION

Stereotype name	Class type	Description	Icon
<< movingobject >>	Class	This stereotype indicates that the class represents a moving object. In our case study the moving object is the mobile hospital	
<< trajectory >>	Class	This stereotype indicates that the class represents a trajectory	
<<trajectorysection >>	Class	This stereotype indicates that the class represents a part of trajectory "trajectory section" that is a component of the trajectory	
<< stop >>	Class	This stereotype indicates that the class represents the "stop" which is a component of the trajectory	
<< move >>	Class	This stereotype indicates that the class represents the	




		movement "move" which is a component of the trajectory	
--	--	--	--

At the end of this section, we propose the following example of defining an UML Class related to the Trajectory Data Class Diagram with an extended stereotype and icon using XML:

```
<STEREOTYPE>
<NAME>movingobject</NAME>
<DESCRIPTION></DESCRIPTION>
<BASECLASSES>
<BASECLASS>UMLClass</BASECLASS>
</BASECLASSES>
<ICON minWidth="30" minHeight="20">mh.bmp</ICON>
<SMALLICON>mh.bmp</SMALLICON>
<RELATEDTAGDEFINITIONSET>movingobject</RELATEDTAGDEFINITIONSET>
```

2) *Association' Stereotypes*: For associations we kept the standard elements of UML such as the association, the generalization, the aggregation and the composition because we noticed that these relations meet users' needs. Besides, we added specific associations that can exist between different components of trajectories as described in the following table:

TABLE III
ASSOCIATIONS STEREOTYPES DESCRIPTION AND REPRESENTATION

Stereotype name	Class type	Description	Icon
<< Inside>>	Association	In our case study the class stop is Inside a given spatial dimension called location	
<< Adjacency>>	Association	Two adjacent trajectory sections share the same stop.	
<<Cross >>	Association	A moving object cross a trajectory or a trajectory section.	

At the end of this section, we propose the following example of defining an UML Association related to the Trajectory Data Class Diagram with an extended stereotype and icon using XML:

```
<STEREOTYPE>
<NAME>Inside</NAME>
<DESCRIPTION></DESCRIPTION>
<BASECLASSES>
```

```
<BASECLASS>UMLAssociation</BASECLASS>
</BASECLASSES>
<ICON minWidth="30" minHeight="20">Inside.bmp</ICON>
<SMALLICON>Inside.bmp</SMALLICON>
<RELATEDTAGDEFINITIONSET>Inside</RELATEDTAGDEFINITIONSET>
</STEREOTYPE>
```

V. TRAJECTORY UML PROFILE REALIZATION

To implement our approach we chose the StarUML open source platform that uses the language XML to create the profiles UML. In this section we describe StarUML by showing its stretchable parts, and then we model a trajectory and their components with our Trajectory-UML profile.

B. The StarUML platform

StarUML is a modeling platform with the UML language, conceived to support the MDA (Model Driven Architecture) approach. It is characterized by a strong flexibility and an excellent extensibility of its features. Indeed, besides the predefined functions, StarUML allows the addition of new functions which can be adapted to the user's needs. The inconveniences of this platform are that it does not allow specifying more than a stereotype for an element and it excludes the definition of the constraints. Thus in our work we considered that every element has only a single stereotype.

C. The implementation of Trajectory-UML profile

An UML profile is one package belonging to the mechanism of extension. This package is stereotypical << Profile >> which is written in XML as we see in the following figure:

```
<? xml version="1.0" encoding="UTF-8" ?>
<PROFILE version="1.0">
<HEADER>
<NAME>TDW</NAME>
<DISPLAYNAME>Data Modeling</DISPLAYNAME>
<DESCRIPTION>TDW Data Modeling Profile</DESCRIPTION>
<AUTOINCLUDE>True</AUTOINCLUDE>
</HEADER>
```

In the StarUML platform, we added an approach named "The tdw Framework" and a profile UML called "TDW model" that contains two diagrams which are respectively; "Trajectory Data Class Diagram" and "Trajectory Data Warehouse Diagram". Indeed, we have created two files XML one for the approach and the other one for the profile. Inside these files we appealed to extensions of notation which are files written in Scheme language (Dialect of LISP) which allows realizing specific notations that are different from those contained in UML.

In this part, we represent the interfaces of our added approach to the platform and the various realized diagrams. When we start StarUML, the dialog box "New Project by approach" appears to choose the wished approach. Below we find our approach called "tdw Framework".

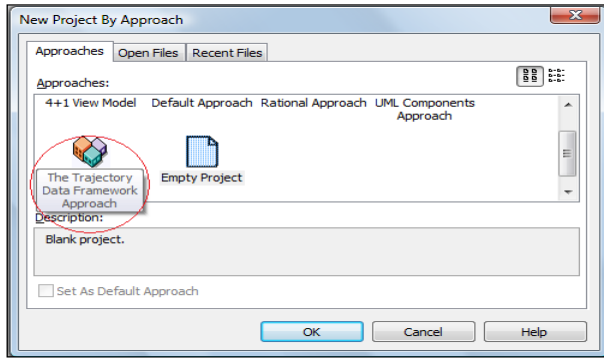


Fig 2. Trajectory data warehouse UML approach

To add diagrams to these models we click the straight button of the mouse, we choose "Add Diagram" and we find the diagrams of our approach. The following figure shows how this step takes place:

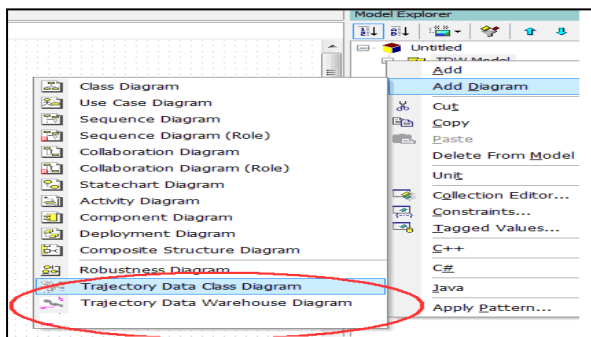


Fig 3. Diagrams of TDW-UML

For every type of diagram we added a palette. These latter allows to visualize the stereotypes and icons of each diagram and to use them. In our case, we created three palettes. The first one is related to the Trajectory Data Warehouse Diagram, the second one is related to the Trajectory Data Class Diagram.

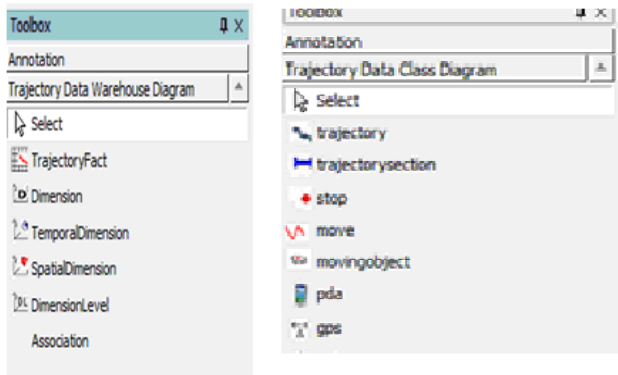


Fig 4. The added pallets

D. Validation of trajectory-UML functioning

Trajectory-UML is guided by some objectives such as the possibility to describe explicitly relationships between trajectories and its components (trajectory-section, stops and moves). Those relationships can be of different types such as topologic, metric, aggregation. Besides, Trajectory-UML allows modeling classical data by offering a well known set of

concepts such as class, attribute, association, generalization, composition. From the ergonomic point of view, the trajectory and its components are visualized in diagrams by pictograms and stereotypes. This allows an immediate unambiguous apprehension of additional features. In this section we model the concept of trajectories and their components with our Trajectory-UML profile.

We created a new diagram called Trajectory Data Class Diagram, in which we added some pictograms and some stereotypes to identify each class (entity). To do this, we used some pictograms of MADS project [23]. The same idea was done in [20] but in a relational model.

In this diagram, there are some stereotypes and icons that can be used in any application related to moving objects. In fact, each moving object has a trajectory. This latter is composed of trajectory sections that are composed of moves and stops. Those latter are in a given location. For a generic Trajectory Data Class Diagram, we propose to keep the classes: Moving object, Trajectory, Trajectory-section, Move, Stop and Location. In the following, we propose the Trajectory Data Class Diagram with Trajectory-UML.

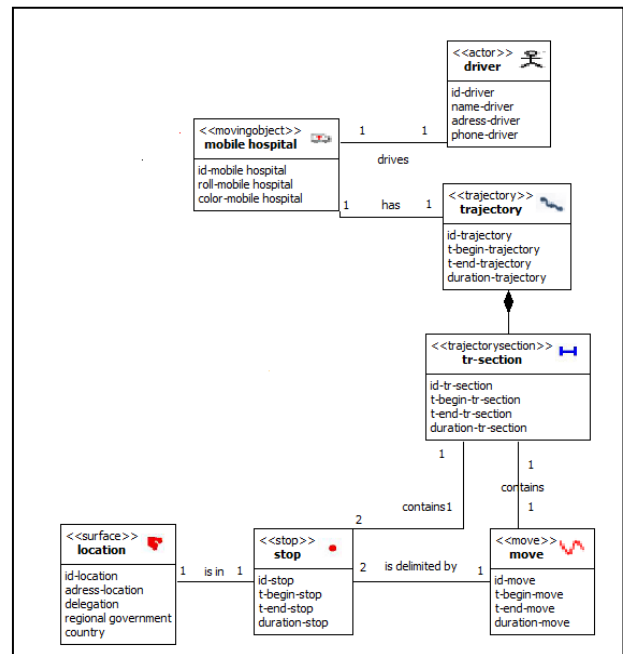


Fig 5. The trajectory data class diagram with our TDW profile

VI. CONCLUSION

In this paper, we described our profile named Trajectory-UML. This profile contains the diagram "Trajectory Data Class Diagram" which gives a conceptual representation of the trajectory of a moving object by specifying relationships between different entities of the diagram. We described the realization of the Trajectory-UML profile. To estimate our approach we ended this paper with an experimentation of the trajectory data class diagram. We propose as future work to represent a model of the component diagram that is based on an UML profile for the physical level.

REFERENCES

- [1] E. wayane,. Four ways to build a data warehouse. 2003.
- [2] T. Karayannidis, N. Sellis, T. MAC: Conceptual data modeling for OLAP, Proc of the International Workshop on DMDW. 2001.
- [3] M, Golfarelli,. From user requirements to conceptual design in data warehouse design. Information science reference, 2009.
- [4] M. Rizzi, S,. WAND: A case tool for workload based design of a data mart. In Proc SEBD, Portoferraio, Italy, pp. 422-426, 2002.
- [5] D. Moody, M. Kortink,. From enterprise models to dimensional models: A methodology for data warehouse and data mart design. In Proc.2nd DMDW,Stockholm, Sweden, 2000.
- [6] J. Ghozzi, O. Teste,. Méthode de conception d'une base multidimensionnelle contrainte, liere journée francophone sur les entrepôts de données et l'analyse en ligne, toulouse, pp. 51-70, 2005.
- [7] Bonifati, A. Cattaneo, F. Designing data marts for data warehouses. ACM transactions on software engineering and methodology, pp. 452-483, 2001.
- [8] Lujan-Mora, S., & Trujillo, J. (2002). Extending UML for Multidimensional Modeling. Alicante University, Spain.
- [9] Lujan-Mora, S., Trujillo, J., Song, Y. (2002). Multidimensional Modeling with UML Package Diagrams.
- [10] S. Lujàn-Mora, J. Trujillo and I. Yeol Song. "A UML profile for multidimensional modelling in data warehouses". Data & Knowledge Engineering (DKE), pp. 725-769. 2006.
- [11] G. Wagner, Eindhoven. "A UML Profile for Agent-Oriented Modeling". Eindhoven Faculty of Technology Management, <http://tmitwww.tm.tue.nl/staff/gwagner>. 2002.
- [12] V. Grassi, R. Mirandola and A. Sabetta. "A UML Profile to Model Mobile Systems". Universit' a di Roma "Tor Vergata", Italy. 2004.
- [13] J. Zubcoff and J. Trujillo. "A UML 2.0 profile to design Association Rule mining models in the multidimensional conceptual modeling of data warehouses". Data & Knowledge Engineering, pp. 44-62. 2007.
- [14] J. Zubcoff, J. Pardillo and J. Trujillo "A UML profile for the conceptual modelling of data-mining with time-series in data warehouses", Information and Software Technology, pp. 977-992 Contents lists, 2009.
- [15] J .Brodeur, Y .Bédard, M.J. Proulx. Modelling Geospatial Application Databases using UML-based Repositories Aligned with International Standards in Geomatics. In Proc. of the ACM Symposium on Advances in Geographic Information Systems, ACM GIS, 2000.
- [16] N. Tryfona, F. Busborg et J.G. Christiansen. "starER: A Conceptual Model for Data Warehouse Design". Dans: Proceedings of the ACM 2nd International Workshop on Data warehousing and OLAP (DOLAP'99), Kansas City, Missouri, USA, pp. 3-8. 1999.
- [17] C. Sapia, M. Blaschka, G. Höfling, B. Dinter. "Extending the E/R Model for the multidimensional paradigm". International Workshop on Data Warehousing and Data Mining (DWDM '98) in conjunction with the 17th Int. Conf. on Conceptual Modeling (ER '98), (Singapore). 1998.
- [18] J. Norberto, M. Juan Trujillo. "An MDA approach for the development of data warehouses ". Decision Support Systems, pp. 41-58. 2008.
- [19] N. Prat, J. Akoka and I. Comyn-Wattiaun. "A UML-based data warehouse design method". Decision Support Systems, pp. 1449-1473. 2006.
- [20] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, C. Vangenot. A Conceptual View on Trajectories. Research Report. Ecole Polytechnique Fédérale, Database Laboratory, Lausanne, Switzerland. May, 29th, 2007.
- [21] Object Management Group (OMG). "Unified Modeling Language Specification 1.5". <<http://www.omg.org/cgi-bin/doc7formal/03-03-01>> . 2003.
- [22] J. Warmer et A. Kleppe. "The Object Constraint Language. Precise Modeling with UML". Object Technology Series, Addison-Wesley. 1998.
- [23] C. Parent, S. Spaccapietra and E. Zimanyi. Conceptual Modeling for Traditional and Spatio-Temporal Applications - The MADS Approach. In Springer Verlag, 2006.

Optimal itinerary planning for mobile multiple agents in WSN

Mostefa BENDJIMA¹
STIC Laboratory
University of Tlemcen
Tlemcen, Algeria

Mohamed FEHAM²
STIC Laboratory
University of Tlemcen
Tlemcen, Algeria

Abstract—Multi mobile agents based on data collection and aggregations have proved their effective in wireless sensor networks (WSN). However, the way in which these agents are deployed must be intelligently planned and should be adapted to the characteristics of wireless sensor networks. While most existing studies are based on the algorithms of itinerary planning for multiple agents i.e. determining the number of mobile agents (MAs) to be deployed, how to group source nodes for each MA, attributing the itinerary of each MA for its source nodes. These algorithms determine the efficiency or effectiveness of aggregation. This paper aims at presenting the drawbacks of these approaches in large-scale network, and proposes a method for grouping network nodes to determine the optimal itinerary based on multi agents using a minimum amount of energy with efficiency of aggregation in a minimum time. Our approach is based on the principle of visiting central location (VCL). The simulation experiments show that our approach is more efficient than other approaches in terms of the amount of energy consumed in relation with the efficiency of aggregation in a minimum time.

Keywords—Wireless sensor network; mobile agent; optimal itinerary; energy efficiency.

I. INTRODUCTION

A wireless sensor network (WSN) consists of a set of nodes capable of collecting information from a monitored environment and for transmission to a base station (Sink) via the wireless medium. WSNs are often characterized by dense deployment in large-scale environments that are limited in terms of resources. The limitations are the insufficient processing, storage capacities and especially energy shortage because they are usually powered by batteries. Constraint on the size of a sensor node requires designers to reduce the size of the battery and therefore the amount of energy available. Replacing a battery is rarely possible, for reasons of cost or constraints due to the environment. This leads to problems related to energy consumption during the operation of different network nodes.

Therefore, unlike traditional networks concerned with ensuring a good quality of service, WSNs must give equal importance to energy conservation. It is widely recognized that energy limitation is an unavoidable issue in the design of WSNs because it imposes strict constraints on the operation of the network. In fact, energy consumption of sensor nodes plays an important role in the life of the network, and has become the predominant criterion of performance in this area.

If we want the network to function satisfactorily as long as possible, these energy constraints force us to compromise between different activities, both at the node and network levels.

Several research studies have emerged with a goal: to optimize energy consumption of nodes through the use of innovative conservation techniques to improve network performance, including the maximization of its life. In general, energy economy remains the best compromise between the different energy-consuming activities. WSNs literature recognizes that the data is a prominent consumer of energy; due to this, the majority of work envisaged techniques affecting this section.

One technique for minimizing energy consumption in sensor networks, which are proposed by the authors such as [1] [2] [3] is the Mobile Agents (MAs) technique. In these proposals, it is crucial to find an optimal itinerary for a mobile agent to perform data collection from multiple distributed sensors.

Therefore, the MA is defined as a message that contains application code, a list of source node predefined by a Sink, and an empty field to put the data [4]. Among the disadvantages of this type of solution is the difficulty of creating areas that will be addressed by the MA. However, as a solution to optimize the itinerary of the MA in data fusion, [5] proposed two heuristic algorithms. A randomly selected itinerary may even cause worse performance than the traditional Client/Server model. For the Local Closest First (LCF) algorithm, the MA itinerary starts from a node and searches for the next destination with the shortest distance to its location. Concerning the Global Closest First (GCF) algorithm, the MA starts its itinerary from a node and selects the next destination closest to the center of the surveillance zone.

In addition, using only a single MA for sensor networks on a large scale can have the following drawbacks:

- A great delay with a large number of source nodes to visit;
- Sensor nodes in the itinerary of the MA deplete energy faster than other nodes;
- During the visit to source nodes, the size increases continuously MA;

- MA transmission will consume more energy in its itinerary back to the sink;
- The increasing amount of data accumulated by the MA during its migration increases its chances of being lost due to the noise in the wireless medium;
- Thus, the longer the itinerary, the more risky is the agent-based migration.

To overcome this problem, several agents can be used, which gives rise to a new problem; itinerary planning for multi agents, which has a significant impact on the performance of the sensor network, knowing that it is crucial to find an optimal itinerary for each MA to visit all indicated source nodes.

Researchers have proposed to partition the network into groups such as cluster algorithms or visit central location with a special node as a Cluster-Head or center for each group respectively. This supports the data exchange with the base station, receiving sensed data from all nodes in one group to send to the Sink.

In this paper, we present an approach adopted to solve the problem of itinerary planning for multiple agents to visit source nodes in each group in parallel. Iteratively, our algorithm partitions the network into sectors where source nodes in each sector are included in an itinerary. Source nodes in each sector can be obtained by choosing the angle in an adaptive manner.

The rest of the paper is organized as follows: Section 2 Provides an overview of the literature in which algorithms for a number of studies have been conducted for itinerary planning of mobile agents in the WSN, first we present the solutions by using a single mobile agent, then the solutions proposed for the use of multi mobile agents. Next, in section 3, we present the problem statement. Our proposed method for partitioning the network into sectors, with the Sink as a starting point, to solve the problem of itinerary planning for multiple agents to visit source nodes in each sector in parallel, is described in section 4. Then, Section 5 sets forth the purpose of our application, which is to establish a system to simulate the communication between a set of sensors and a base station forming a wireless sensor network. Finally, Section 6 summarizes and concludes this paper.

II. STATE OF THE ART

A number of studies have been conducted for mobile agent itinerary planning in the sensor network. For itinerary planning, first we present the solutions proposed by the use of a single mobile agent, then we focus on the solutions proposed for the use of multi mobile agents.

A. The use mobile agent

Itinerary planning of a single mobile agent has a significant impact on the effectiveness of sensor networks. It determines the order of visits to source nodes during mobile agent migration. Among these proposals, Local Closest First (LCF) and Global Closest First (GCF) are simple approaches to itinerary planning of a single mobile agent [6]. The LCF algorithm seeks the source node with the shortest distance from the current node, while the GCF algorithm searches for

the following source node, which has the shortest distance from the Sink. The author Chen proposed Mobile-Agent-Directed Diffusion (MADD) algorithm in [7] that is similar to the LCF, but chooses the farthest source node as the starting point of the itinerary.

There are LCF algorithms that extend by estimating the cost of communication in the migration of the MA, with the increase of the MA packet size. Among these algorithms [8], Itinerary Energy Minimum for First source selection (IEMF) and Itinerary Energy Minimum Algorithm (IEMA) have the highest performance in terms of energy efficiency and delay compared to existing solutions. To address this performance, a simplified analytical model is used to formulate the itinerary calculated by an objective function directly proportional to the received signal strength and inversely proportional to data loss and energy consumption.

Restrictions on the use of a single agent to perform the entire task make the algorithm unreliable in applications where a large number of sources nodes are to be visited.

Typically, itinerary algorithms with a single agent have a high efficiency when the source nodes are distributed geographically close to each other, and the number of source nodes is not high.

For sensor networks on a large scale, in which many sources nodes have to be visited, the itinerary of a single mobile agent presents the following disadvantages:

- Significant delays when a single agent has visited hundreds of sensor nodes;
- Sensor nodes in the itinerary of the mobile agent deplete energy faster than other nodes;
- The size of the mobile agent increases when visiting more source nodes, so during transmission, the agent will consume more energy in its itinerary back to the base station;
- Reliability is reduced when the agent accumulates an increasing amount of data;
- When the mobile agent migrates to several source nodes, the chance of being lost increases.

To address these problems, multi mobile agents can be used, which causes a new problem requiring itinerary planning for multiple agents.

B. The use of multiple mobile agents

Chen *et al.* [9] proposed the algorithm Centre Location-based Multi agents Itinerary Planning (CL-MIP), the main idea is to consider the solution of multi agent itinerary planning (MIP) as an iterative version of the solution the single mobile agent itinerary planning (SIP).

Contributions of the CL-MIP include the following:

- It first proposes a generic framework in several steps to solve the MIP problem, which reuses existing solutions in SIP;
- The proposed gravity algorithm accurately describes the density center of source node groups, which is the basis of the source group.

Another algorithm, Genetic Algorithm-based Multi agents Itinerary Planning (GA-MIP) is proposed in [10]. GA-MIP first proposes a new method for two-level encoding of MA to solve the MIP problem. The GA-MIP algorithm considers the problem as a single problem instead of using several steps, as adopted by the previous algorithms.

The proposal in [11] considers models of MIP problems as a Totally Connected Graph (TCG). In the TCG, the vertices are the nodes of the sensor network, and the weight of an edge is derived from estimates of the jump between the two end nodes of the edge. The authors indicate that all source nodes in a particular sub-tree should be considered as a group. In addition, the authors present a balancing factor while calculating the weight in the TCG, to form a minimum Balanced Spanning Tree (BST). The balancing factor allows flexible control over the compromise between energy cost and duration of the task. This algorithm is named Balanced minimum Spanning Tree-based Multi agents Itinerary Planning (BST-MIP). The main contribution of BST-MIP is that it analyzes the critical impact of geographical positions on the source node group.

Mpitiopoulos et al. in [12] proposed a Near-Optimal Itinerary Design (NOID) algorithm to solve the problem of computing a near-optimal itinerary for MA. NOID algorithm was designed so as to adapt rapidly to changing conditions of the network, the MA itinerary must contain only source nodes with enough energy availability and the number of AMs should depend on the number of physical locations of the source nodes to be visited.

Gavalas et al. in [13] presented the second quasi-optimal itinerary design algorithm (SNOID) to determine the number of mobile agents that should be used, and the itinerary of these mobile agents should follow. The main idea behind SNOID is to partition the area around the Sink into concentric zones and begin to build paths for MA with the direction of the inside near Sink. The radius of the first area that includes Sink equals (a.rmax) where a is an input parameter in the range [0, 1] and rmax is the maximum transmission range of any source node. All source nodes inside the first zone are connected directly to Sink, and are the starting points of mobile agent itinerary.

III. PROBLEM STATEMENT

A. VCL algorithm

The author of [1] proposed selection algorithm of the Visiting Centre Local (VCL), it assumes that we have sensor nodes n, so this algorithm is based on the principle of distributing the impact factor of each sensor node to other sensor nodes. Then, each sensor node will receive (n-1) impact factors from other sensor nodes, and of itself. After calculating the accumulated impact factor, the location of the sensor node with the largest cumulative impact factor is selected as (VCL). We denote the set of n nodes in V.

For two source nodes $i, j \in V$, $d(i, j)$ is the distance between i and j. Then, we can estimate the number of hops between these two source nodes as: $H_{ij}^1 = \frac{d(i,j)}{R}$ where R represents the maximum transmission range. To approximate the effects of a real gravitational field, a Gaussian function is adopted to

calculate the impact factor between two source nodes i and j:

$$G_{ij} = e^{-\frac{(H_{ij}^1 - 1)^2}{2\sigma^2}}, \text{ where } \sigma \text{ is a constant, set to } 08 \text{ in all previous works.}$$

B. LCF algorithm

We chose the algorithm Local Closest First (LCF) to trace the itinerary of the MA as it allows finding an optimal way to avoid redundancies, and to manipulate lists of nodes in a consistent manner. According to [5] the itinerary starts from the local point of a sub-area, our strategy aims requires the Sink to be the starting points for the packet itinerary.

IV. DESCRIPTION OF THE APPROACH

The main purpose of our strategy is to partition the nodes of wireless sensor network into groups. The partition is done iteratively so that source nodes are included in an itinerary. The data, collected by the source nodes in each group, are arranged in a group. Then, the data of each group of source nodes are sent as a pack instead sending them separately by each source nodes to the Sink.

The main idea is to partition the network into sectors, with the starting point is the base station. For each sector, we propose a mobile agent to collect information from these source nodes along an itinerary similar in shape to the sector.

Once the Sink receives signals from source nodes, it will send mobile agents, so that they circulate between its nodes using the itinerary Local Closest First algorithm (LCF). The mobile agents aggregate data processed and collected by the nodes to return to the Sink with the information collected.

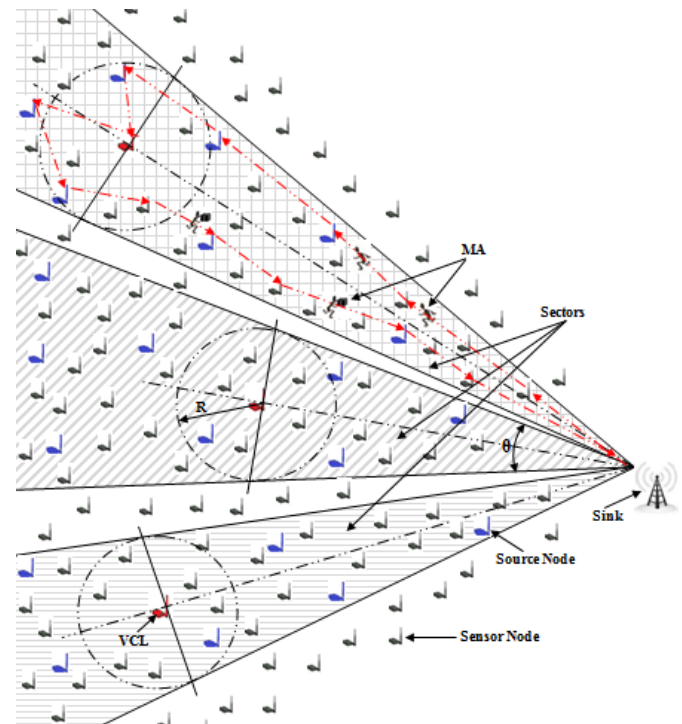


Figure 1. Source nodes partitioning into sectors

As shown in “Fig.1”, the disk radius is an important parameter in source nodes grouping. Similarly, the selection of the angle θ determines the size of the sector, which is important for an efficient itinerary. Data aggregation of MAs by the source nodes from all sectors passes through the following steps:

- calculating the VCL source nodes;
- Constructing an angle sector θ and a right-side center between the Sink and VCL;
- Grouping the source nodes inside the sector;
- We repeat these steps for all the remaining source nodes;

After that, all source nodes are sources included into the sectors, a mobile agent is assigned to each sector. At this point, the itinerary of the MA can be planned for each sector, using the LCF algorithm.

All these steps are translated in three algorithms using the following notation:

NOTATION:

VCL: visiting center local;

θ : angle of the sector;

R: maximum transmission range;

S: sector for source node grouping;

T: set of source nodes to be grouped in S sectorateachiteration;

N: total number of nodes of sources;

E: set of source nodes;

E': set of remaining source nodes;

SN: source node.

ALGORITHM 01: ALGORITHM FOR PARTITION INTO SECTORS

BEGIN

- *Plotting a straight line connecting the Sink and the VCL;*
- *Plotting a straight line perpendicular to the line connecting the Sink and the VCL (Sink, VCL), the VCL is the point of intersection;*
- *Plotting a circle (disk) whose center is the VCL and with a radius R;*
- *Connecting the two points of intersection of the straight line perpendicular and the circle with the Sink;*

End.

Algorithm 02: algorithm to group nodes of sources

Begin

$T = \{\};$

For each source node SN in E do

If $SN \in S$ then

$T = T \cup \{SN\};$

$E = E - SN;$

END IF;

END FOR;

END.

ALGORITHM 03: ALGORITHM FOR ITINERARY PLANNING MMA (MULTI MOBILE AGENT)

Begin

While $E \neq \{\}$ do

Selection algorithm VCL // (see VCL algorithm [14]);

Partitioning algorithm into sectors // (see Algorithm 01);

Algorithm to group source nodes // (see Algorithm 02);

$E = E - T;$

End while;

Determine the itinerary of AM in each sector area by LCF (see LCF algorithm [05]);

End.

V. SIMULATION SETUP

A. Purpose of the software

For our simulation platform, we have established a system to simulate communication between a set of sensors and a base station constituting a wireless sensor network. The technique we used consists in partitioning the network into sectors, and using mobile agents, in order to save the network energy. So the goal of our application is to carry out itinerary planning of mobile multi agents that is effective in terms of energy and in terms of task duration.

B. Initial hypotheses

To perform the simulation, we adopted the following assumptions:

- At the beginning, each sensor has an initial energy;
- We can choose the number of nodes, the radius and position of Sink;
- The number of sensors is limited by a minimum and a maximum value;
- The positions of the sensor nodes are known to the Sink;
- We can add sensor nodes before the activation of the nodes only;
- The deployment is random;
- Two sensors do not occupy the same coordinates;
- We consider that the problem of failure is caused by a depletion of energy;
- The simulation stops when a target cannot be covered by at least one sensor;
- The aggregation model and energy used is mentioned in reference [15].

We performed our simulations on a 1000m x 500m with a random distribution of 1000 sensor nodes for 1000 seconds. Thus, the base station is located to the right of the field, and multiple source nodes are randomly distributed in the network. We have limited the radio coverage, and brought the data rate of each node to 80 meters and 1Mbps, respectively, as suggested in [16]. The parameters of power transmission and reception, which directly affect the radio range, were selected from the ranges defined in the sunspot system [17]. Local processing time is 40 ms. The parameters used for our simulation are summarized in Table I. The values we used are inspired by the work done by [18].

TABLE I: BASIC SIMULATION PARAMETERS

Simulation parameter	Values
Node distribution	Random
Radio range	80m
Debit	1Mbps
Data collected interval	10 seconds
Simulation time	1000 seconds
Local processing time	40ms
Processing code size	0.4Ko
Raw Data Size	2024 bits
MA Code Size	1024 bits
MA Accessing Delay	10 ms
Data processing Rate	50 Mbps
Raw Data Reduction Ratio	0.8
Aggregation Ratio	0.9
Network Size	1000m x 500m
Number of Sensor Nodes	1000

C. Results and analysis

Itinerary planning for a single mobile agent allows determining the order of source nodes visited during migration of a mobile agent, it has a significant impact on the effectiveness of sensor networks MA-based. Several studies have been conducted for itinerary planning in a single mobile agent in a WSN. Among these works, Local Closest First (LCF) and Global Closest First (GCF) are simple approaches [19] for itinerary planning of a single mobile agent. LCF seeks the source node according to the shortest distance from the current node, while the GCF seeks the source node which has the shortest distance from the base station.

To demonstrate the performance of our Multi Mobile Agent Itinerary (MMAI) approach in wireless sensor networks, we compare it with LCF and GCF. In this section, we examine the impact of the number of source nodes on the energy performance and the duration of tasks. Sensor nodes are energy limited, except for the Sink which is assumed to have infinite energy input. We assume that the Sink and sensor nodes are stationary and that the Sink is located on the right side of the field. To check the scaling property of our algorithms, we select a large-scale network with 1000 nodes, so we set the number of source nodes from 05 to 50 by steps of 05, and get a set of results for each case. We represented the results obtained in "Fig. 2", which shows the impact of the number of nodes on energy sources, for sensory data of all

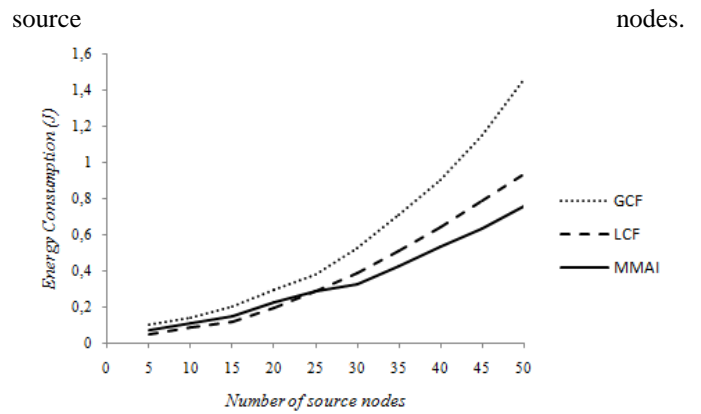


Figure 2. Energy consumption by the number of source nodes

The analysis of the "Fig. 2" highlights several interesting elements: first, which is quite normal, when the number of source nodes is increased, more energy is required for mobile agents to perform tasks for each of the three approaches. However, for an equal number of source nodes, consumption using our MMAI approach is always lower than that of LCF and GCF. Moreover, the difference between our approach and the approach GCF starts digging from 05 source nodes, while the gap with other LCF begins at 25 sources nodes. When the number of nodes is low, energy consumption of our approach is less important compared to LCF, between 05 and just under 25 nodes, the latter achieves an additional energy saving 18% higher than MMAI. It may be noted again that from 25 source nodes the difference between MMAI and the other two approaches is becoming increasingly important, and this difference increases continuously with increasing number of nodes. Nevertheless, from just over 25 nodes, our MMAI approach significantly outperforms the two other approaches; for example, at 30 nodes; LCF and GCF consume 16% of energy and 37% more than MMAI, respectively. However, at 50 source nodes, consumption using MMAI approach is 21% lower than in the LCF approach and 48% less compared to the GCF approach. By comparison, the multi mobile agent itinerary solution has better energy efficiency.

In addition, and in another experiment, we show the performance comparison of the three approaches in terms of task duration. The results are shown in "Fig. 3".

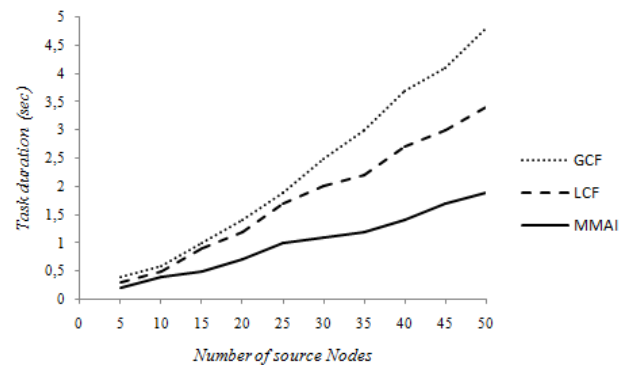


Figure 3. Task duration made by the number of nodes

“Fig. 3” compares the performance of the three approaches in terms of the duration of tasks. We note that in terms of task duration our approach is still shorter than that of other approaches, regardless of the number of source nodes. In the single mobile agent approach as LCF and GCF, the duration of a task starts from the moment the mobile agent is distributed by the base station until the agent returns to the station.

In multi mobile agent approach such as MMAI, since many mobile agents are working in parallel, there must be an MA which is the last to return to the Sink. Then, the duration of the task of the MMAI approach is the delay of that agent.

Compared to the energy performance, the number of nodes sources has a greater impact on the delay of the task. First, when the number of source nodes is increased, a longer period of time is required for mobile agents to perform the duties of each of the three approaches. It may be noted again that the difference between our proposal and the other two approaches is becoming increasingly important, and this difference increases continuously with the increasing number of source nodes. The duration of a task in an approach with a single mobile agent as LCF and GCF becomes much greater for a greater number of source nodes, because with more source nodes to visit, the size of an MA becomes bigger, and more transmissions are made.

At the beginning, the three approaches are almost similar, but at 25 source nodes, LCF and GCF consume 41% and 48% more time than MMAI, respectively. Our approach is advantageous with increasing source nodes, nodes with 50 sources, MMAI minimize delay, over 44% in LCF, and approximately 61% in GCF.

The reason for this result is that in the itinerary of a single MA, only one MA moves along the entire network to collect information from all source nodes. This procedure leads to a greater latency from source nodes that can be distributed throughout the network. A multi agents approach can speed up the task because several itineraries are applied simultaneously. As Figure 03 shows, the duration of the task of the proposed MMAI is still the lowest in comparison with the other two approaches.

VI. CONCLUSION AND PERSPECTIVES

In this paper, we proposed a solution to the problem of itinerary planning for multiple in wireless sensors networks. Recent research has demonstrated the effectiveness of using a single mobile agent for data collection in wireless sensor networks. Using a single mobile agent may be deficient, for example, the increased latency in task completion and unbalanced energy consumption by the nodes. Nevertheless, it is crucial to find an optimal itinerary for each MA to visit all the source nodes. To remedy this problem, we partitioned the network into sectors, and within each sector, we use a mobile agent to collect information from these source nodes along an itinerary with a shape similar to the sector. Therefore, we used multi mobile agents to visit the source nodes in each group in parallel, which facilitates the simultaneous collection of sensory data to reduce the duration of tasks in depth. Iteratively, our algorithm partitions the network into areas

where the nodes in each sector sources are included in an itinerary; the starting point is the base station. Once the Sink receives signals from source nodes, it will send mobile agents, so that they circulate between its nodes using the algorithm groups LCF. These mobile agents perform data aggregation of processed and collected data by the source nodes to return to Sink with the information collected.

Many simulations have been performed to show the superior performance of our MMAI approach compared to LCF and GCF approaches in terms of the duration of tasks and energy consumption. It saves significant energy and thus significantly reduces costs on a large scale.

For future work, we will attempt to improve the selection method of the angles of the sector, so that the optimal itinerary for MAs can be obtained by selecting the effective angle in an adaptive mode.

REFERENCES

- [1] M. Chen, T. Kwon, Y. Yuan, and V. C. M. Leung, “Mobile Agent Based Wireless Sensor Networks,” pp. 14–21, 2006.
- [2] T. Lang, Z. Qing, and A. Srihari, “Sensor networks with mobile agents,” Multidisciplinary University Research Initiative (MURI) under the Office of Naval Research and Army Research Laboratory CTA on Communication and Network under Grant DAADIP-OI-2-CQII, 2003.
- [3] Q. Wang, and H. Xu, “Mobile-agent-based collaborative signal and information processing in sensor networks,” Proceedings of the IEEE 91(8), 2003.
- [4] M. Ketel, S. Numan, and A. H. Dogan, “Distributed Sensor Networks based on Mobile Agents Paradigm,” Dept. of Electrical & Computer Eng. North Carolina A&T State University Greensboro, NC 27411. IEEE, 2005.
- [5] Q. Hairong, and W. Feiyi, “Optimal Itinerary Analysis for Mobile Agents in Ad Hoc Wireless Sensor Networks,” University of Tennessee, Knoxville, Advanced Networking Group MCNC, Research Triangle Park, 2001.
- [6] F. Mourchid, “Nouveau Modèle pour le Positionnement des Senseurs avec Contraintes de Localisation,” memoire the master applied sciences, AVRIL 2010.
- [7] M. Chen, “Mobile Agent-Based Directed Diffusion in Wireless Sensor Networks,” EURASIP J. Advances in Processing, vol. 2007.
- [8] M. Chen, T. Laurence, T. Yang, and Z. Liang, “Itinerary Planning for Energy-Efficient Agent Communications in Wireless Sensor Networks,” IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, VOL. 60, NO. 7, SEPTEMBER 2011.
- [9] M. Chen, S. Gonzalez, Y. Zhang, and V.C. Leung, “Multi-agent itinerary planning for sensor networks,” Proc. IEEE 2009 Int. Conf. Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine 2009), Las Palmas de Gran Canaria, Spain, 2009.
- [10] W. Cai, M. Chen, T. Hara, and L. Shu, “GA-MIP: genetic algorithm based multiple mobile agents itinerary planning in wireless sensor network,” Proc. Fifth Int. Wireless Internet Conf. (WICON), Singapore, 2010.
- [11] M. Chen, W. Cai, S. Gonzalez, and V.C. Leung, “Balanced itinerary planning for multiple mobile agents in wireless sensor networks. Proc. Second Int. Conf. Ad Hoc Networks (ADHOCNETS2010), Victoria, BC, Canada, 2010.
- [12] D. Gavalas, A. Mpitziopoulos, G. Pantziou, and C. Konstantopoulos, “An approach for near-optimal distributed data fusion in wireless sensor networks,” Springer Wirel. 16, (5), pp. 1407–1425, 2009.
- [13] P. Sameera, P. Sundeep, K. Bhaskar, and S. Gaurav, “A unifying framework for tunable topology control in sensor,” University of Southern California, USA, 2005.
- [14] M. Chen, V. Leung, S. Mao, and T. Kwon, “Energy-efficient Itinerary Planning for Mobile Agents in Wireless Sensor Networks,” In: IEEE. 2007. International Conference on Communications, Dresden, Germany, 2009.

- [15] C. Konstantopoulos, A. Mpiziopoulos, D. Gavalas, and G. Pantziou, "Effective determination of mobile agent itineraries for data aggregation on sensor networks," *IEEE Trans. Knowl. Data Eng.* pp. 1679–1693, 2010.
- [16] K. Sohrawy, and D. Minoli, "Wireless Sensor Networks, Technology, Protocols, and Applications," WILEY, 2007.
- [17] SunTM, Small Programmable Object Technology (SunSPOT) Theory of Operation. Technical report, Sun Microsystem, SunLabs, 2008.
- [18] A.sardouk, and I. boulahia, "agent-cooperation based communication architecture for wireless sensor network," in: *ifip/ieee wireless days/adhoc and wireless sensor networks, uae, ifip*, 2008.
- [19] M. chen, s. gonzlez, y. zhang, and v. leung, "multiagent itinerary planning for sensor networks," in *proceedings of the ieee 2009 international conference on heterogeneous networking for quality, reliability, security and robustness (qshine 2009)*, laspalmas de gran canaria, spain, 2009.

Genetic procedure for the Single Straddle Carrier Routing Problem

Khaled MILI

Applied Economics and Simulation,
High institute of management
2000 Bardo, Tunisia

Faissal MILI

Applied Economics and Simulation,
Faculty of Management and Economic Sciences,
5100 Mahdia, Tunisia

Abstract— This paper concentrates on minimizing the total travel time of the Straddle Carrier during the loading operations of outbound containers in a seaport container terminal. Genetic Algorithm is well-known meta-heuristic approach inspired by the natural evolution of the living organisms. Heuristic procedure is developed to solve this problem by recourse to some genetic operators. A numerical experimentation is carried out in order to evaluate the performance and the efficiency of the solution.

Keywords- Component; Container terminal optimization; routing Straddle Carrier; heuristic; Assignment strategy; Genetic Algorithm.

I. INTRODUCTION

The success of a seaport container terminal resides in a fast transshipment process with reduced costs and it is measured by many performance indicators such as the productivity and the customer satisfaction. Because containerhips are highly capitalized and their operating costs are very high, it is very important that their turn-around time in container terminals is as short as possible. The turn-around time of a containerhip implies the sum of the queue time and service time which is the sum of time for berthing, unloading, loading and departing.

Since inbound containers are usually unloaded into a designated open space, the yard handling equipment (i.e. Straddle Carrier) does not have to travel much during the unloading operation. However, the time for loading depends on the loading sequence of containers as well as the number of loaded containers. The Single Straddle Carrier Routing Problem (SSCRP) has a great effect on the time service and the performance of the container terminal. In this paper, our objective is to minimize the total travel time of the Straddle Carrier (SC) for loading outbound containers.

The remainder of the paper is organized as follows: section 2 present the related works that solve the SSCR. Section 3 will be reserved to detail the problem formulation. The two following sections will be reserved to present our Genetic Algorithm solution and our paper will be finished by numerical examples in order to prove the efficiency of our method. Some concluding remarks and perspectives to extend this work are finally discussed.

II. RELATED WORKS

Operational decision problems on seaport terminals have received increasing attention by researchers. Some of them evoked the SCRP as Kim and Kim ([1] [2] [3]); they present in

their papers a mathematical formulation for the SSCR where their objective is to minimize the distance between yard-bays in the storage area during the SC tour. They propose a solution procedure based on beam search heuristic method that its principle is to select the nearest yard-bay to visit by a SC.

However, they suggest that the times spend by the SC inside yard bays are constant. In reality, we note that these times vary and significantly affect the SC routing tour. In fact, we judge that the reshuffling and the unproductive movements inside a yard bay depend on the positions of required containers into stacks and levels. Hence, we will evaluate the time spent between initial position and all destinations yard bays in addition to the time cost inside each one and we select the least of them. Therefore, in our paper we solve the SSCR by considering this time (intra-yard bay) as variable.

Little research has been done on this topic although the practical importance of this problem ([4] [5] [6] [7] [8]).

In 2003, V. Nunes Leal Franqueira [9] proposed Heuristic strategies Beam Search and Ant Colony Optimization to solve the single vehicle routing problem. These are tested comparatively. A new strategy for container collection is proposed as a substitute for the traditional greedy strategy of container collection. The proposed strategy increased the number of alternatives considered within the search space and turned out to improve the quality of solutions. It is proved that the Ant colony heuristic react well as beam search method in this strategy.

In 2005, E. Nishimura et al. [10] present in their paper a Genetic Algorithm heuristic to solve the trailer routing problem using a dynamic routing assignment method. They focus on the tours related to one cycle operation of the quay cranes. Experimental results demonstrate that the dynamic assignment is better than static one. The drawback to their solution procedure is the complexity of the trailer routing, which may increase the possibility of human error. Trailer drives may find difficult to follow the complicated itineraries assigned to them, resulting in mistakes in driving.

E. Nishimura et al. in 2001 [11] address in their paper the problem of a dynamic assignment ships to berths. They develop a heuristic procedure based on genetic algorithm where chromosomes are presented as character strings instead of bit binary string representation. The chromosome representation of the solution defines the set assignments of

ships to berths, giving ships the positions of service sequence for each berth.

III. PROBLEM FORMULATION

A. Straddle Carrier definition

By a “subtour” of a SC, we mean a visiting sequence of yard-bays which a SC visits to pick up all the containers which will be loaded onto a cluster of cells in the ship. An overview of a container terminal is presented in Figure 1.

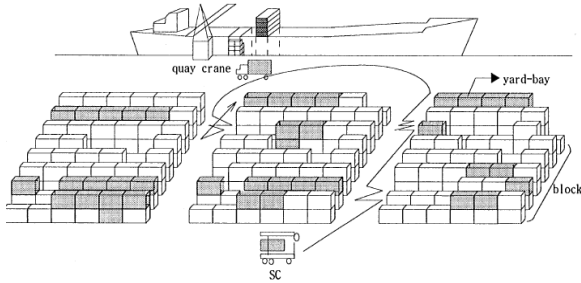


Figure 1. An overview of a container terminal

B. Optimization model

An optimization model will be developed to display the container arrivals and yard locations and the actual and optimized assignment of straddles to containers.

The main part of this modeling is to develop and evaluate the algorithms for assigning straddles to containers.

The discussion from the previous section illustrates the fact that the manner in which the straddles are assigned container jobs impacts the cost and service quality of operation.

In general, the problem of assigning straddles to containers can be formulated as the assignment problem, a mathematical programming problem presented by L.N. Spasovic et al. in 1999 [12].

$$\text{Min } Z = \sum_i \sum_j c_{ij} x_{ij} \quad (1)$$

s.t

$$\sum_{i=1}^n x_{ij} = 1 \quad \text{for } j = 1, 2, \dots, n \quad (2)$$

$$\sum_{j=1}^n x_{ij} = 1 \quad \text{for } i = 1, 2, \dots, n \quad (3)$$

Where

$i, j =$ indices

$n =$ number of containers.

$$x_{ij} = \begin{cases} 1 & \text{if the feasible assignment of container "i" to straddle "j" is selected} \\ 0 & \text{otherwise} \end{cases}$$

$c_{ij} =$ cost of the (i, j) assignment

Equation (1) is an objective function that minimizes costs. Constraints (2) and (3) are typical assignment problem

restrictions that ensure that a straddle can be assigned to only one container and vice versa.

IV. GENETIC ALGORITHMS (GA)

A. Motivation

Because of existing the several equality constraints in our model, the implementation of the GA in order to quick and facilitate achieve to the feasible solutions is the objective of our work.

Our reasons for choosing GA as a solution approach are as follows:

We need an approach to search the feasible route. The most of the model's constraints are as equality form and, therefore, obtaining of the feasible solutions is a hard task. In this case, the probability of reaching infeasible solutions is more than feasible solutions and therefore we need a population-based approach such as GA to better exploration of the solution routing.

GA is a well-known meta-heuristic that its efficiency is verified for many problems in the literature.

GAs work on a population of the solutions simultaneously. They combine the concept of survival of the fittest with structured, yet randomized, information exchange to form robust exploration and exploitation of the solution routing [13].

B. Probleme outlines

The problem can be summarized as follow:

- Each feasible solution of the problem is treated as a chromosome in the GA.
- Every container must be picked up once and exactly once at any route.
- The handling time of each container is dependent on its positions in the storage yard.

The set of chromosomes construct a generation where steps of GA are applied:

Step1: A fitness function is applied to extract the fittest value of the generation

The SCRП is a minimization problem; thus the smaller the objective function is, the higher the fitness value must be.

Step2: The GA selects the fittest chromosomes and applied a crossover operator to give rise to better solution.

The crossover scheme should be capable of creating new feasible solution (child) by combining good characteristics of both parents.

Step3: These solutions are treated by a mutation operator which introduces random changes to the chromosomes to create new generation

C. Our representation

In our GA application, a candidate solution to an instance of the SCRП specify the number of required containers, the possible visited yard bays, the partition of the demands and

also the delivery order for each route. Each chromosome represents a feasible solution.

We will apply genetic operators to generate new individuals. Each of them defines a possible route of genes that represent the required containers.

For our problem a chromosome is a set of containers that can be visited by a SC to perform a quay crane (QC) work schedule. For example a QC demands r_t containers type A to load them into a containership, SC has to move toward the yard bays that include this type of containers, and transports them to the QC. Each container has a transportation cost depends on its position inside the storage yard and especially in its chromosome's order. The number of genes in a chromosome is bounded by r_t required containers. And the cost of the SC's tour will be the sum of transportation costs of the picked up containers.

The position is defined by the number of yard bay, stack and level. Every type or group is presented by a matrix. Each element of the matrix will be equal to 1 if the group of the current container is the same as the required type and 0 otherwise. Therefore we begin our procedure by this set of required containers from which we construct the initial generation of n chromosomes. The GA will choose randomly a set of chromosomes that contains r_t genes. The gene represents a required container.

Let $C_{s,l}^i$ designs the container inside yard bay i in stack s in level l ; i.e. $C_{8,3}^1$ represents a container inside the first yard bay, in the third level of the 8th stack.

V. GENETIC OPERATORS

A. Fitness function

For our solution procedure we will take under consideration the sigmoid function as defined in E. Nishimura et al. in 2001 [11] ; where $z(y)$ denotes the objective function value.

$$f(y) = 1 / (1 + \exp(z(y) / 10.000)) \quad 0 \leq f(y) \leq 0.5 \quad (4)$$

For the feasible solutions, our GA calculates the cost of each route that satisfies the objective function of the SSCR. Then it compares between these costs and selects the smallest amount one.

B. Reproduction

It is a process in which chromosomes are copied according to their scaled fitness function values, i.e., chromosomes with a higher fitness value would have more of their copies at the next generation. This can be done by randomly selecting and copying chromosomes with probabilities that are proportional to the fitness values (costs of routes presented in each chromosome).

1) *Initial Situation:* We have n genes representing the number of the available required containers.

Example: for $n=9$

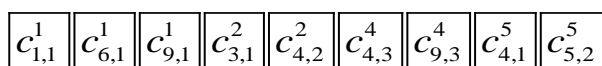


Figure 2. list of required containers

2) *Initial generation:* GA chooses randomly r_t genes from this table, constructs generations of chromosomes and selects the two fittest ones from each generation.

For our example, let chromosomes A and B be the selected parents.

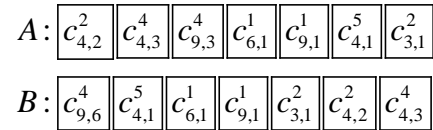


Figure 3. Parent's chromosomes

C. Crossover

We use the 2-point crossover to introduce new chromosomes (or children) by recombining current genes. In this crossover, two cut points are randomly chosen on the parent chromosomes. In order to create an offspring, the strings between these two cut points in both chromosomes are interchanged.

A crossover may generate infeasible children in terms of constraint, i.e., a child chromosome may have container to be picked up twice. In order to keep the feasibility, the crossover operation is performed in the following manner.

First, substrings to be interchanged are given by two crossover points that are randomly determined. After interchanging the substrings between chromosomes A and B , we have chromosomes A' and B' as temporary children.

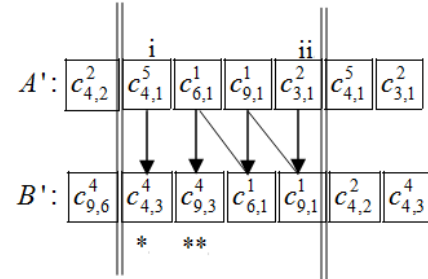


Figure 4. Temporary children

They are infeasible, because for instance, chromosome A' has containers $C_{4,1}^5$ and $C_{3,1}^2$ to be picked up twice and lose containers $C_{4,3}^4$ and $C_{9,3}^4$. Next, additional interchanges described below are carried out to make them feasible. Letting the interchanged string be the substring that was interchanged so far in each chromosome, we examine genes from left to right in the interchanged string of A' . Note that we never interchange any genes in the interchanged string again in the following process:

As the most left gene is container $C_{4,1}^5$ which is scheduled to be picked up twice in A' , obtain the container type (which is $C_{4,3}^4$) in B' at the corresponding position of the gene (hereafter referred to as a cell).

Therefore the container $C_{4,1}^5$ placed in the 7th bit in the chromosome A' will be replaced by $C_{4,3}^4$.

The second double container in chromosome A' is $C_{3,1}^2$.

When we look at the corresponding gene in chromosome B' we find the container $C_{9,1}^1$ which already exists in A' , so we follow the interchange arrows until we reach its end. In our example we stop in the container $C_{9,3}^4$ which will replace $C_{3,1}^2$ in chromosome A' .

We look at chromosome B' and we follow the same way as in A' .

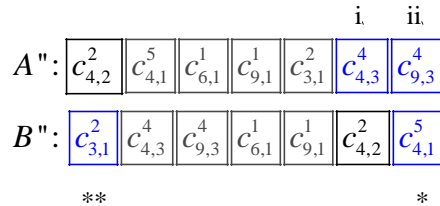


Figure 5. children after crossover

D. Mutation

This genetic operator introduces random changes to the chromosomes by altering the value to a gene with a user-specified probability called mutation rate. As an example we choose to apply the swap operator to each gene. This operator consists in randomly selecting two genes and exchanging them.

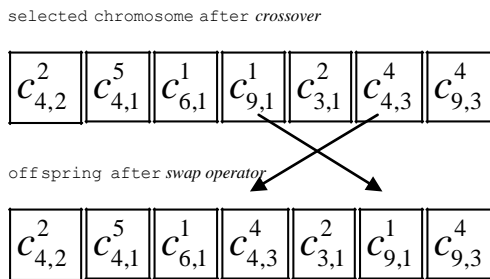


Figure 6. Swap operator

VI. SOLUTION HEURISTIC PROCEDURE

At really quay crane work schedule, many cases can be found. In this paper the ‘single tour’ and the ‘subtours’ cases are chosen.

Case 1: single tour (where containers of the same type are required)

1. We have initial situation with all available required containers
2. Repeat
 - Create a generation in which all chromosomes have r genes where r is equal to the number of required containers.
 - Calculate the objective function for each individual
 - Transform it to a fitness value
 - Select best-ranking individuals to reproduce
 - let individuals having better fitness be new parents
 - Create offspring through crossover operator

- Evaluate the individual fitness of the offspring
 - Select the chromosome having the biggest fitness value.
 - Apply mutation operator on the selected chromosome.
 - Select the chromosome having the best solution
- Until the number of required generations is reached
3. Select the best solution from all resulting chromosomes

Case 2: subtours (many subtours to perform)

Let t design a number of subtour $\{t = 1 \dots T\}$

Example:

($t=1$) first subtour to transport containers type A

($t=2$) second subtour; transport containers type B .

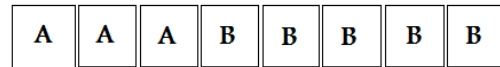


Figure 7. chromosome of two subtours

The genetic procedure for case2 is presented as follow:

Step1: For $t = 1$ (first subtour)

- let r_t containers (genes) from a type h
 - apply genetic algorithm developed in case 1
- Step2: For subtours $t = \{2 \dots T\}$
- start from the last straddle carrier’s position of the resulting chromosome in step1
 - Apply genetic algorithm developed in case 1

VII. NUMERICAL EXAMPLES

The resulting solutions of two strategies (Greedy collection strategy and Random collection strategy) are used as references for our experiments. The Greedy strategy is the collection of the maximum quantity of containers available at a determined location. The Random strategy is the collection of a random quantity of containers without leaving behind containers needed to complete a work schedule subtour. A MATLAB program was used to solve the above mixed-integer programming for an example problem. The MATLAB is a programming environment for algorithm development, data analysis, visualization, and numerical computation. All experiments were performed on a 3GHz Pentium 4 computer.

In the following tables the presented solutions of the GA procedure are the best ones between 12 generations for every iteration.

Case 1: in this case containers of the same group (A) are required. Two instances are generated in order to evaluate the GA procedure.

TABLE I. INSTANCES (CASE1).

	Instance 1	Instance 2
Type of required containers	A	A
Number of required containers	10	70

Number of available containers	66	215
Number of yard bays	6	20
Number of 3 level's stacks	10	15

TABLE II. GENETIC ALGORITHM PROCEDURE (GA) WITH THE GREEDY COLLECTION STRATEGY (GCS) AND RANDOM COLLECTION STRATEGY (RCS): COSTS AND PROCESSING TIMES (IN CASE1).

Instance2	Costs (sec)	GCS	12916				
		RCS	15421	13886	15190	15335	14968
		GA	15003	14372	14817	14834	15173
	Processing times (sec)	32.4	62.9	68.7	70.4	75.1	

By analyzing the results in table II and table IV the following remarks can be made:

- In instance 1, Genetic algorithm demonstrated to return solutions of better quality than the Random collection strategy and the Greedy collection approach, within the same range of the processing time.
- Due to the random nature of Genetic algorithm heuristic method, increasing the size of the problem in instance 2 to look for solutions does not necessarily guarantee that a better solution will be reached. As observed in iterations 5 and 7, Random collection strategy found a better quality solution.
- By looking at iterations 4 and 8 in instance 2, we notice that Greedy collection strategy worked better since it found the minimum costs. It can be explained by the fact that the genetic heuristic selects the next location based on container's positions, i.e. both the yard-bay and the position of containers (stacks and levels) are taken into account, while the Greedy collection strategy selects the next location based on the yard-bays only.
- It seems that increasing the number of subtours has the same effect as increasing the number of the required containers.
- In big size problem, genetic procedure is less performed, this is can be explained by the fact that at each generation built by genetic procedure, a number of selected containers is fixed and the genetic operators are used only in this set of containers in order to choose the best cost solution between them.
- Genetic algorithm demonstrates better solution sin small sizes problem.
- The processing time in all iterations is quite reasonable.

VIII. CONCLUSION

In this paper, we have presented the routing problem of the SC to load outbound containers. We use Genetic Algorithm as an approach to solve the SSCRP, we defined its utility, and we developed our Genetic heuristic procedure and its operators such as fitness function, reproduction, crossover and mutation in order to get best solution to choose a route for the SC which has the least cost. Numerical examples are carried out to prove the performance and the efficiency of our method comparing with greedy and random collection strategies.

IX. FUTURE WORK

In a real container terminal there is more than one quay crane, with their respective work schedules, and many Straddle Carriers (SCs). Several SCs must complete their routing, by sharing the same container yard-map. This new scenario introduces a few complications to the single SC routing problem, increasing significantly the routing problem

		Iterations		1	2	3	4	5
Instance1	Costs (sec)	GC	1138.7					
		S						
		RCS	1138.7	1180.0	1179.3	1138.7	1180.0	
	GA	694.4	747.6	672.8	667.0	695.1		
Processing times (sec)		0.85	0.94	1.08	0.52	0.78		
Instance2	Costs (sec)	GC	8269.8					
		S						
		RCS	7860	8503.1	8980.4	8503.8	7649	
	GA	7835.8	7937.8	7646.3	8819.1	8891.1		
Processing times (sec)		32.4	31.3	33.67	32.13	34.63		
		Iterations		6	7	8	9	10
Instance1	Costs (sec)	GC	1138.7					
		S						
		RCS	967.8	1179.3	1140.9	1179.3	1138.7	
	GA	684.4	798.0	680.7	721.8	741.8		
Processing times (sec)		0.85	0.55	0.75	1.33	1.34		
Instance2	Costs (sec)	GC	8269.8					
		S						
		RCS	8912.6	7845.6	8579.7	8093.1	8473.2	
	GA	8265.9	8254.6	8293.1	8959.8	8263.5		
Processing times (sec)		32.4	32.15	33.48	31.54	32.44		

Case 2: in this case containers of the different groups are required. Two instances are generated in order to evaluate the GA procedure.

TABLE III. INSTANCES (CASE2).

	Instance 3		Instance 4	
Type of required containers	A	B	A	B
Number of required containers	10	8	75	60
Number of available containers	66	68	215	240
Number of yard bays	6		20	
Number of 3 level's stacks	10		15	

TABLE IV. GENETIC ALGORITHM PROCEDURE (GA) WITH THE GREEDY COLLECTION STRATEGY (GCS) AND RANDOM COLLECTION STRATEGY (RCS): COSTS AND PROCESSING TIMES (IN CASE2).

		Iterations		1	2	3	4	5
Instance1	Costs (sec)	GCS	1659.9					
		RCS	2168.3	1733.3	2202.4	1944.8	1999.3	
		GA	1770.1	1398.9	1486.6	1698.1	1633.4	
	Processing times (sec)		0.85	2.5	2.42	0.39	0.45	
Instance2	Costs (sec)	GCS	12916					
		RCS	14175	14115	13396	15249	14691	
		GA	13931	17668	17051	14391	16931	
	Processing times (sec)		32.4	75.0	69.5	59.8	72.1	
		Iterations		6	7	8	9	10
Instance1	Costs (sec)	GCS	1659.9					
		RCS	1944.8	1684.8	1946.3	1724.6	1800.6	
		GA	1839.9	1786.0	1417.6	1734.9	1703.8	
	Processing times (sec)		0.85	0.5	1.51	2.30	1.95	

complexity. Our future work aims to solve the multiple straddle carrier routing problem (MSCRP) in a container terminal. In Other metaheuristics can be applied in the future, such Ant colony and particle swarm optimization, and comparative study can be done.

REFERENCES

- [1] K. H. Kim, K. Y. Kim, A routing algorithm for a single transfer crane to load export containers onto a containership , Computers & Industrial Engineering 33 (1997) 673- 676.
- [2] K. H. Kim, K. Y. Kim, Routing straddle carriers for the loading operation of containers using a beam search algorithm, Computers & Industrial Engineering 36 (1999) 109-136.
- [3] K. H. Kim, K. Y. Kim, A routing algorithm for a single straddle carrier to load export containers onto a containership, Int. J. Production Economics 59 (1999) 425- 433.
- [4] DW. Cho, Development of a methodology for containership load planning. Ph.D. thesis, Oregon State University, 1982.
- [5] YG. Chung, SU. Randhawa, ED. McDowell, *A simulation analysis for a transtainer-based container handling facility*, Computers Ind Eng 14(2) (1988) 113±25. [
- [6] F. Soriguera, D. Espinet and F.Robuste, A simulation model for straddle carrier operational assessment in a marine container terminal, Journal of Maritime Research 2006.
- [7] K. M. Van Hee, R.J. Wijbrands, *Decision support system for container terminal planning*, European Journal of Operational Research 34 (1988) 262–272.
- [8] C. Zhang, J. Liu, Y.-W. Wan, K.G. Murty, R.J. Linn, Storage space allocation in container terminals, Transportation Research. Part B: Methodological 37 (2003) 883– 903.
- [9] V. Nunes Leal Franqueira, Single Vehicle Routing in Port Container Terminals, Master thesis, Universidade Federal do Espirito Santo, Brazil. August 2003.
- [10] E. Nishimura, A. Imai, S. Papadimitriou, *Yard trailer routing at a maritime container terminal*, Transportation Research Part E: Logistics and Transportation Review, Volume 41, Issue 1, (2005) 53-76.
- [11] E. Nishimura, A. Imai, S. Papadimitriou, Berth allocation in the public berth system by genetic algorithms, European Journal of Operational Research 131 (2001) 282– 292.
- [12] L N. Spasovic, et all, “Increasing productivity and service quality of the straddle carrier opérations at container port terminal”, Journal of Advanced Transportation October 20, 1999.
- [13] M. Bazzazi, , N. Safaei, , N. Javadian. “A genetic algorithm to solve the storage space allocation problem in a container terminal.” Computers & Industrial Engineering, vol. 56, pp. 44-52, April 2009.

AUTHORS PROFILE

Khaled MILI

Ph. Doctor in quantitative methods,

A member of Laboratory of operational research (LARODEC),

High institute of management

E-Mail: khaledmili@yahoo.fr

Phone: 00216 52 181 176

Faissal MILI

Ph. Doctor in quantitative methods,

A member of Applied Economics and Simulation laboratory,

Faculty of Management and Economic Sciences, 5100 Mahdia, Tunisia

E-Mail: milisoft@yahoo.fr

Phone: 0021620545728

The Virtual Enterprise Network based on IPSec VPN Solutions and Management

Sebastian Marius Rosu, Marius Marian Popescu
Radio Communications & IT Department
Special Telecommunications Service
Bucharest, Romania

George Dragoi, Ioana Raluca Guica
Department of Engineering in Foreign Languages
University "Politehnica" of Bucharest
Bucharest, Romania

Abstract—Informational society construction can't be realized without research and investment projects in Information and Communication Technologies (ICT). In the 21st century, all enterprises have a local area network, a virtual private network, an Intranet and Internet, servers and workstations for operations, administration and management working together for the same objective: profits. Internet Protocol Security is a framework of open standards for ensuring private communications over public networks. It has become the most common network layer security control, typically used to create a virtual private network. Network management represents the activities, methods, procedures, and tools (software and hardware) that pertain to the operation, administration, maintenance, and provisioning of networked systems. Therefore, this work analyses the network architecture for an enterprise (industrial holding) geographically dispersed. In addition, the paper presents a network management solution for a large enterprise using open source software implemented in the PREMINV Research Center, at the University "Politehnica" of Bucharest.

Keywords—Enterprise network; IPSec; network architecture; VPN; network management.

I. INTRODUCTION

Under the concept of a global economy, enterprises are assigning design and production environments around the world in different areas. The requirement for highly reliable and available services has been continuously increasing in many domains for the last decade [1]. The optimization of product benefit must be the focus of all network activities [2]. The work comprises components for integration of information systems, visualization of the planning and production situation, communication to enable cooperative decision making under uncertainty, optimization of plans and simulation of the decisions, network diagnostics and performance monitoring among others [3]. This involves a number of challenges such as providing members access to network-wide real time information, enable visualization of the available information, secure the interaction between advanced information and communication technologies (ICT) based decision support tools and human decision making, creating a coordinated and collaborative environment [2] for planning and decision making. The implementation phase in ICT system is done by doing several socialization activities such as training, hands-on workshop, coaching or even giving a grant for the users who use the system correctly [4].

Monitoring of such process execution may allow the manager to detect faults and guarantee correct execution [5] - e. g. voicedata packets have to arrive at the destination in time, with a defined cadence and with low and constant delay in order to allow the real time voice reconstruction [6]. Because the new communication system enables many more interactions between many more participants, it has security requirements beyond the conventional confidentiality, integrity and availability properties provided by conventional security systems [7]. The idea of NGN (Next Generation Network) is developed with the purpose of integrating different multiple services (data, voice, video, etc.) and of facilitating the convergence of fixed and mobile networks [8]. However, the effects of ICT devices on the productivity of companies cannot be measured unequivocally at the microeconomic level because of certain statistical and methodological imperfections, the difficulties in measuring network effect at a business level and the lack of data enabling to make international comparisons [9]. Development of information technology and communication has led to widespread deployment of technical solutions for [10]:

- Accessing and processing data and information;
- The transmission of data and information in a network environment with distributed destinations;
- Connect different users regardless of their geographical distance and position.

The complexity of the human enterprise continues to grow at an accelerating pace as larger numbers of people take on increasingly ambitious tasks in a world that grows in size, complexity, and constraining factors [11]. On-line applications (e.g. e-banking, electronic voting, information sharing and searching) require anonymous measures to prevent third parties from gathering online private information. As a general requirement for an infrastructure support is that the enterprises must be able to inter-operate and exchange information and knowledge in real time so that they can work as a single integrated unit [12], although keeping their independence/autonomy.

Various network services can be used by everyone, either supplying or demanding them. A large range of distribution, the platform independence, a big number of user friendly services that are easily accessible through the World Wide Web as well as the open standards used and free or budget-priced products (such as browsers, html editors, software

updates) have lead to a high and continuously growing proliferation of the Internet [13]. More and more people are using Internet to access information and communicate with each other. Development of ICT leaves much more freedom to the designers and consultants to accommodate organizations to other influences, both internal and external [14]. Enterprises are now facing growing global competition and the continual success in the marketplace depends very much on how efficient and effective the companies are able to respond to customer demands [15]. Starting from these considerations, this work analyzed the virtual enterprise network (VEN) architecture for an enterprise geographic dispersed as support for virtual private networks (VPNs) possible structures (based on Internet Protocol Security – IPSec) and presents a network monitoring solution using open source software to enterprise business improvement.

II. THE ENTERPRISE NETWORK GENERAL ARCHITECTURE

An enterprise network consists of a group (departmental, interdepartmental, etc.) of *local area networks* (LANs), located in the same place or geographically dispersed, interconnected using *wide area networks* (WANs) and contains a number of inter-networking devices (e.g. switches, routers, gateways, etc.) which is under the control of the organization or a telecommunication company. A communication network forms the backbone of any successful organization [16]. Metropolitan networks play a critical role in the overall expansion of network services because they not only provide for services within individual metropolitan areas, but they also serve as the gateways for wide-area national - and international - scale networks [17]. In an enterprise network, a large number of nodes are interconnected together through a computer network as follow [18]:

- End-user nodes represented by access points such as workstations, personal computers, printers, mainframe computers, etc.
- Network active elements consist of devices such as multiplexers, hubs, switches, routers, and gateways; the active elements and links provide the needed physical communication paths between every pair of end-user nodes.

Today, traditional infrastructures type Internet/Intranet/Extranet have now a fast dynamic, marking the transition to new generation networks to provide higher speeds to the user (end to end), for different types of activities and transactions and a significant reduction in the number of servers by passing information between two nodes [19].

In the last decade, specially, the idea of virtual enterprise (VE) called on a *virtual enterprise network* (VEN) or a *virtual enterprise business network* (VEBN) is meant to establish a dynamic structure of the organization by a synergetic combination of dissimilar enterprises (i.e. small and medium sized enterprises) with different core competencies, thereby forming a best of everything temporary alliance in an industrial group or holding to perform a given business project to achieve maximum degree of customers requirements and customer satisfaction [15].

In the last years, the trend toward IP-based transport infrastructures for all real-time and non-real-time applications opens the door for a new paradigm in integrated voice and data communications. A hierarchical network design model breaks the complex problem of network design into smaller, more manageable problems [15]. An important step in designing an enterprise network is to define a network perimeter. The enterprise network perimeter (see figure 1) defines a security layer complemented with other security mechanism [20], [21]. Communications within and outside the enterprise perimeter must be through a traffic control point - provided by firewalls and other security devices [19]. Large area networks (WANs, specific for large enterprise or for businesses geographically dispersed) were designed to solve connection problems between different workstations and different local networks, or only a local network where the distances are too large to be able to use a simple cable connection. The network designs are examples of secure network architecture that are scalable for home offices, small and medium sized businesses, or business enterprises. A variety of hardware, operating systems, and applications can be used in their implementation. Both commercial and free open source products can be used for the workstations, web servers, security servers, and database servers.

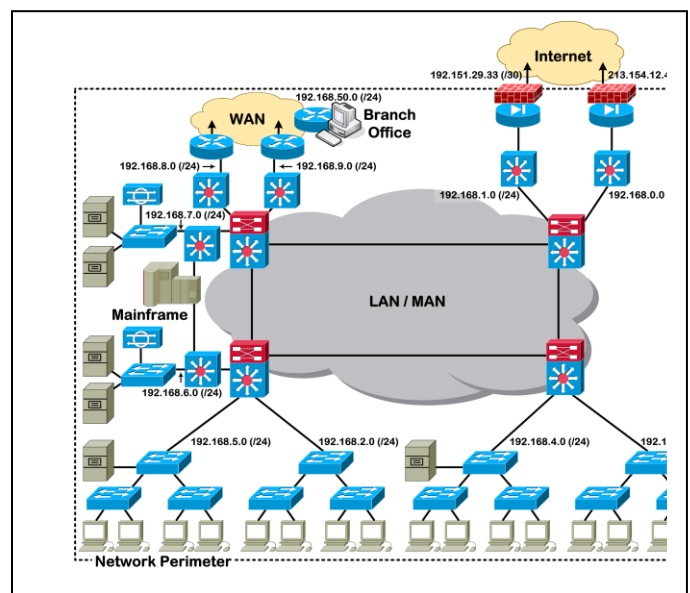


Figure 1. The enterprise network perimeter

A high performance backbone has an intrinsic value for an ultra-fast Internet connection only if the points of connection and network users, POP (*Point of Presence*), providing an equivalent level of performance. ATM (*Asynchronous Transfer Mode*) is a packet-switched technology that uses virtual circuits over a single physical connection from each location to the ATM cloud. Data is transferred in cells or packets of a fixed size. The small, constant cell size allows ATM equipment to transmit video, audio and computer data over the same network. ATM creates a fixed channel between two points whenever data transfer begins. This makes it easier to track and bill data usage across an ATM network (see figure 2), but it makes it less adaptable to sudden surges in network traffic.

Four types of service are available: *constant bit rate* (CBR), *variable bit rate* (VBR), *available bit rate* (ABR) and *unspecified bit rate* (UBR). In this idea, we propose in figure 3 a general virtual enterprise network architecture for a large

enterprise or an industrial holding (with headquarters and branches) formed by a temporary alliance of different small and medium sized enterprises, geographically dispersed, with ATM Points of Presence (PoPs).

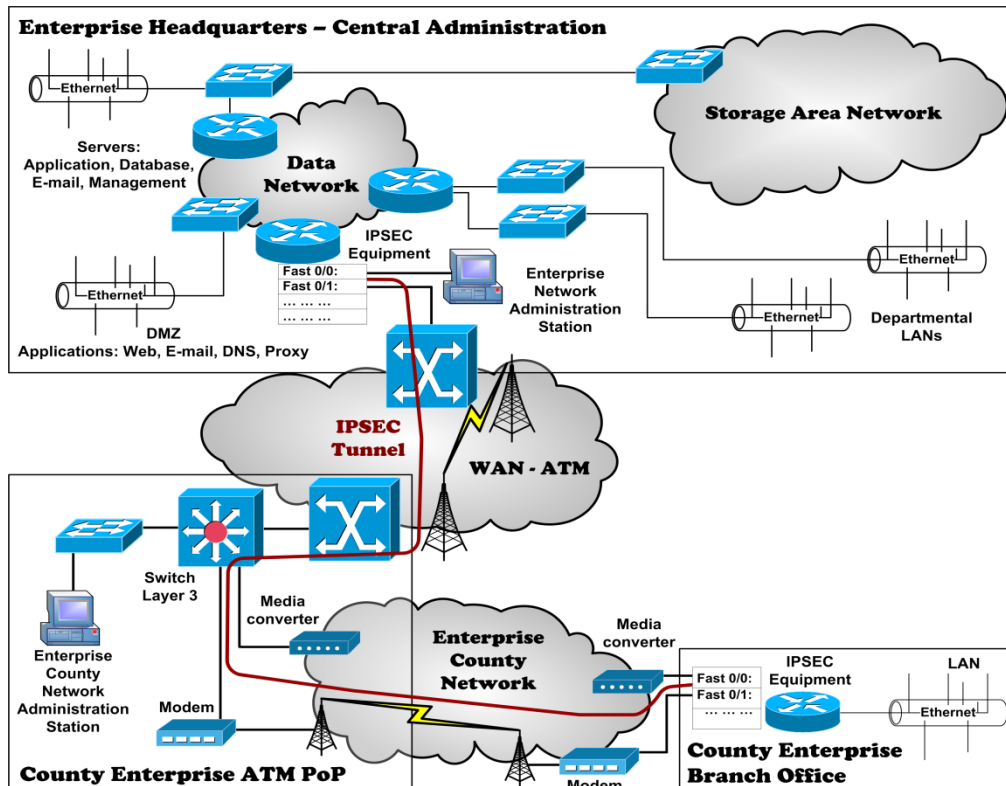


Figure 2. An ATM network for an enterprise geographically dispersed with PoPs to support transfer of large volumes of data over long distances

This solution is implemented in the PREMINV e-platform at the University “Politehnica” of Bucharest, where the virtual enterprise network (VEN) is based on a virtual private network (VPN). This is an emulated network built on public infrastructure (shared), and particularly dedicated to a client (the private) to connect the different users in locations and capable to ensure similar conditions of integrity, confidentiality and quality similar with those of a private network.

Virtual Enterprise Networks allows the provisioning (i.e. private network services) for a dynamic organization over a public or shared infrastructure such as the Internet or service provider backbone network. Appearance of a virtual enterprise network is related to the evolution switches. The first and the most important role of a virtual enterprise network is to realize a synergetic combination of a group of users regardless of their geographical position but in such a manner that it flows together and provide the best performances. Secondly, a VEN provide administrative solutions which accompany the products, allowing users moving from one group to another through a simple reconfiguration of the equipment [22]. However, the application-to-application communication problem still exists. Businesses have needed a standardized way for applications to communicate with one another over networks; no matter how those applications were originally implemented [23].

III. THE ENTERPRISE VPNS IPSEC SOLUTIONS

In fact, emulated VPN is a network build on public infrastructure (shared), dedicated to a client (privacy), to connect the users and to ensure the conditions of integrity, confidentiality and quality similar with a private network. It purposes the following classification of VPN's:

- a) After the length of structures:
 - Permanent VPN
 - Enabled VPN (tunneling): Client Tunnel Compulsory Tunnel.
- b) As responsible for implementing:
 - VPN's provider's responsibility
 - VPN's client responsibility
 - VPN's provider's and customer responsibility

If VPN is the responsibility of the supplier and is reduced to connectivity, the content and inter-location communication are the responsibility of the recipient (customer) and the provider should not restrict the type of inter-location intercommunication that only it would to the extend that it would have repercussions on the physical network).

- c) Type of access in VPN:
 - VPN remote access (Dial to client);

- Intranet VPN (site to site model);
 - Extranet VPN (Business to Business model).
- d) After expanding territories of:
- Local VPN;
 - National VPN;
 - International VPN.
- e) After topology:
- Hub and Spoke VPN;
 - Any to any VPN;
 - Hybrid VPN.
- f) After the type of date:
- VPN “pure” (connectivity);
 - VPN and contents services (content) (Internet, voice, voice VPN, video) ;
 - Content provider is the provider of VPN;
 - Content provider other than VPN provider.

- *Host-to-gateway* – protects communications between one or more individual hosts belonging to a specific network of an organization. Host-to-gateway is used to allow hosts of unsecured networks, access to internal organization services such as email and web servers.
- *Gateway-to-gateway* – this model protects communications between two specific networks, such as organization’s headquarters networks and organization’s branch offices or two business partners’ networks.

IPSec is a framework of open standards for ensuring private communications over public networks. It has become the most common network layer security control, typically used to create a virtual private network (VPN). IPSec Tunnel mode is used to secure gateway-to-gateway traffic. IPSec Tunnel mode is used when the final destination of the data packet is different from the security termination point. IPSec Tunnel mode protects the entire contents of the tunneled packets. The IPSec Tunnel mode data packets sent from the source device are accepted by the security gateway (a router or a server) and forwarded to the other end of the tunnel, where the original packets are extracted and then forwarded to their final destination device [26]. IPSec tunnel is usually built to connect two or more remote LANs via Internet so that hosts in different remote LANs are able to communicate with each other as if they are all in the same LAN. Common commands to create an IPSec tunnel (for Cisco® equipments) are presented in figure 4 (connects Enterprise headquarter LAN through an IPSec tunnel to 2 Enterprise Branch Office LANs). A VPN solution based on IPsec (see figure 5) typically requires integration of several services (design, network management services, dial-up or dedicated access).

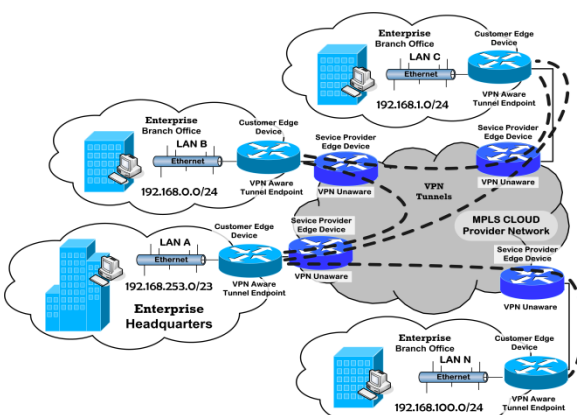
Figure 3. VPN IP/MPLS Network

VPN provisioned using technologies such as Frame Relay and Asynchronous Transfer Mode (ATM) virtual circuits (VC) have been available for a long time, but over the past few years IP and IP/Multi-Protocol Label Switching (MPLS) – based VPNs have become more and more popular (see figure 3). The central idea of MPLS is to attach a short fixed-length label to packets at the ingress router of the MPLS domain.

Packet forwarding then depends on the tagged label, not on longest address match, as in traditional IP forwarding [24]. VPN may be service provider or customer provisioned and falls into one of two broad categories: site-to-site VPN which connect the geographically dispersed sites of organizations and remote access VPN which connect mobile or home-based users to an industrial holding [25]. There are three primary models for VPN architectures that can be implemented at the enterprise level [15]:

- *Host-to-host* – used to protect communication between two computers. The model is most used when a small number of users must be online or is given a remote that requires protocols that are normally uncertain.

A VPN solution typically requires integration of several services (design, network management services, dial-up or dedicated access). The trend is now evolving to intranets and extranets defined logic, which will lead to the reintegration of the various networks in a single logical subdivision. Structures that allow the approximation of this goal are virtual private networks. Possible solutions in the PREMINV platform to implement a VPN structure for a VE system realization in a geographically dispersed enterprise (see figure 5) can be [15] [26]: local VPN based on VLAN (Virtual Local Area Network), local VPN based on IPSec (Internet Protocol Security), VPN wide area based on IPSec, VPN wide area based on MPLS (Multi-Protocol Label Switching), VPN based on PPPoL2TP (Point-to-Point Protocol over Layer 2 Tunneling Protocol), etc.



Enterprise Headquarters Cisco® 1811 Router	Enterprise Branch Office 1 Cisco® 831 Router	Enterprise Branch Office 2 Cisco® 831 Router
<pre>sh run Building configuration... Current configuration : 8527 bytes ! version 12.4 service timestamps debug datetime msec service timestamps log datetime msec service password-encryption ! hostname ENTERPRISE ! boot-start-marker boot-end-marker ! enable secret 5 \$1\$rP5K\$... .. ! no aaa new-model ! ip cef ! no ip domain lookup ! username cisco privilege 8 password 7 01100F175804 ! crypto isakmp policy 1 authentication pre-share group 2 crypto isakmp key NRjI... address 193.151.29.2 crypto isakmp key MJr3... address 193.151.30.2 crypto isakmp keepalive 10 ! crypto ipsec transform-set SET esp-des esp-sha-hmac ! crypto map BRANCH 10 ipsec-isakmp set peer 193.151.29.2 set transform-set mirades match address 101 ! crypto map BRANCH 20 ipsec-isakmp set peer 193.151.30.2 set transform-set mirades match address 102 ! interface FastEthernet0 description LINK_to_L3 ip address 193.151.31.2 255.255.255.248 ip virtual-reassembly duplex auto speed auto crypto map BRANCH ! interface FastEthernet1 no ip address shutdown duplex auto speed auto ! interface FastEthernet2 ! interface FastEthernet3 ! interface FastEthernet4 ! interface FastEthernet5 ! interface FastEthernet6 ! interface FastEthernet7 ! interface FastEthernet8 ! interface FastEthernet9 ! interface Vlan1 description LAN ip address 192.168.157.1 0.0.0.7 ! ip route 0.0.0.0 0.0.0.0 193.151.31.1 ! no ip http server no ip http secure-server ! access-list 101 permit ip 192.168.157.0 0.0.0.7 192.168.57.0 0.0.0.7 access-list 102 permit ip 192.168.157.0 0.0.0.7 192.168.57.8 0.0.0.7 ! control-plane ! privilege exec level 8 traceroute privilege exec level 8 ping privilege exec level 8 show configuration privilege exec level 8 show ! line con 0 line 1 modem InOut stopbits 1 speed 115200 flowcontrol hardware line aux 0 line vty 0 4 password 7 045802150C2E login local transport input telnet ssh ! end</pre>	<pre>sh running-config Building configuration... Current configuration : 2132 bytes ! version 12.4 no service pad service timestamps debug datetime msec service timestamps log datetime msec no service password-encryption ! hostname BRANCH_OFFICE_1 ! boot-start-marker boot-end-marker ! enable secret 5 \$1\$HdL9\$... .. ! no aaa new-model ! dot11 syslog ! ip cef ! username cisco privilege 8 secret 5 \$1\$TdMn\$LuvKyj7ZHw8rm8Pz7DIsm/ ! crypto isakmp policy 1 authentication pre-share group 2 crypto isakmp key GYz... address 193.151.31.2 crypto isakmp keepalive 10 ! crypto ipsec transform-set SET esp-des esp-sha-hmac ! crypto map BRANCH 10 ipsec-isakmp set peer 193.151.31.2 set transform-set mirades match address 101 ! archive log config hidekeys ! interface FastEthernet0 ! interface FastEthernet1 ! interface FastEthernet2 ! interface FastEthernet3 ! interface FastEthernet4 description WAN ip address 193.151.29.2 255.255.255.248 duplex auto speed auto crypto map BRANCH ! interface Vlan1 description LAN ip address 192.168.57.1 255.255.255.248 ! ip forward-protocol nd ip route 0.0.0.0 0.0.0.0 193.151.29.1 ! no ip http server no ip http secure-server ! access-list 101 permit ip 192.168.57.0 0.0.0.7 192.168.157.0 0.0.0.7 ! control-plane ! privilege exec level 8 traceroute privilege exec level 8 ping privilege exec level 8 show crypto isakmp sa privilege exec level 8 show crypto isakmp privilege exec level 8 show crypto privilege exec level 8 show configuration privilege exec level 8 show ! line con 0 no modem enable line aux 0 line vty 0 4 login local ! scheduler max-task-time 5000 end</pre>	<pre>sh running-config Building configuration... Current configuration : 2132 bytes ! version 12.4 no service pad service timestamps debug datetime msec service timestamps log datetime msec no service password-encryption ! hostname BRANCH_OFFICE_2 ! boot-start-marker boot-end-marker ! enable secret 5 \$1\$HdL9\$... .. ! no aaa new-model ! dot11 syslog ! ip cef ! username cisco privilege 8 secret 5 \$1\$TdMn\$LuvKyj7ZHw8rm8Pz7DIsm/ ! crypto isakmp policy 1 authentication pre-share group 2 crypto isakmp key KFj... address 193.151.31.2 crypto isakmp keepalive 10 ! crypto ipsec transform-set SET esp-des esp-sha-hmac ! crypto map BRANCH 10 ipsec-isakmp set peer 193.151.31.2 set transform-set mirades match address 101 ! archive log config hidekeys ! interface FastEthernet0 ! interface FastEthernet1 ! interface FastEthernet2 ! interface FastEthernet3 ! interface FastEthernet4 description WAN ip address 193.151.30.2 255.255.255.248 duplex auto speed auto crypto map BRANCH ! interface Vlan1 description LAN ip address 192.168.57.9 255.255.255.248 ! ip forward-protocol nd ip route 0.0.0.0 0.0.0.0 193.151.30.1 ! no ip http server no ip http secure-server ! access-list 101 permit ip 192.168.57.8 0.0.0.7 192.168.157.0 0.0.0.7 ! control-plane ! privilege exec level 8 traceroute privilege exec level 8 ping privilege exec level 8 show crypto isakmp sa privilege exec level 8 show crypto isakmp privilege exec level 8 show crypto privilege exec level 8 show configuration privilege exec level 8 show ! line con 0 no modem enable line aux 0 line vty 0 4 login local ! scheduler max-task-time 5000 end</pre>

Figure 4. VPN IPSec tunnel configuration between the enterprise headquarters and branch offices using Cisco® equipments

Newer, VPN can be used in different ways to support business processes. This is the ideal solution if it is not efficient in terms of construction costs a particular network for a firm with a workforce highly mobile, or for small firms that can not justify the cost of their telecommunications network.

Also, VPN can be purchased from a telecommunications company and as an alternative they can use existing network infrastructure as the Internet or public switched telephone network and software through the tunnel crossing.

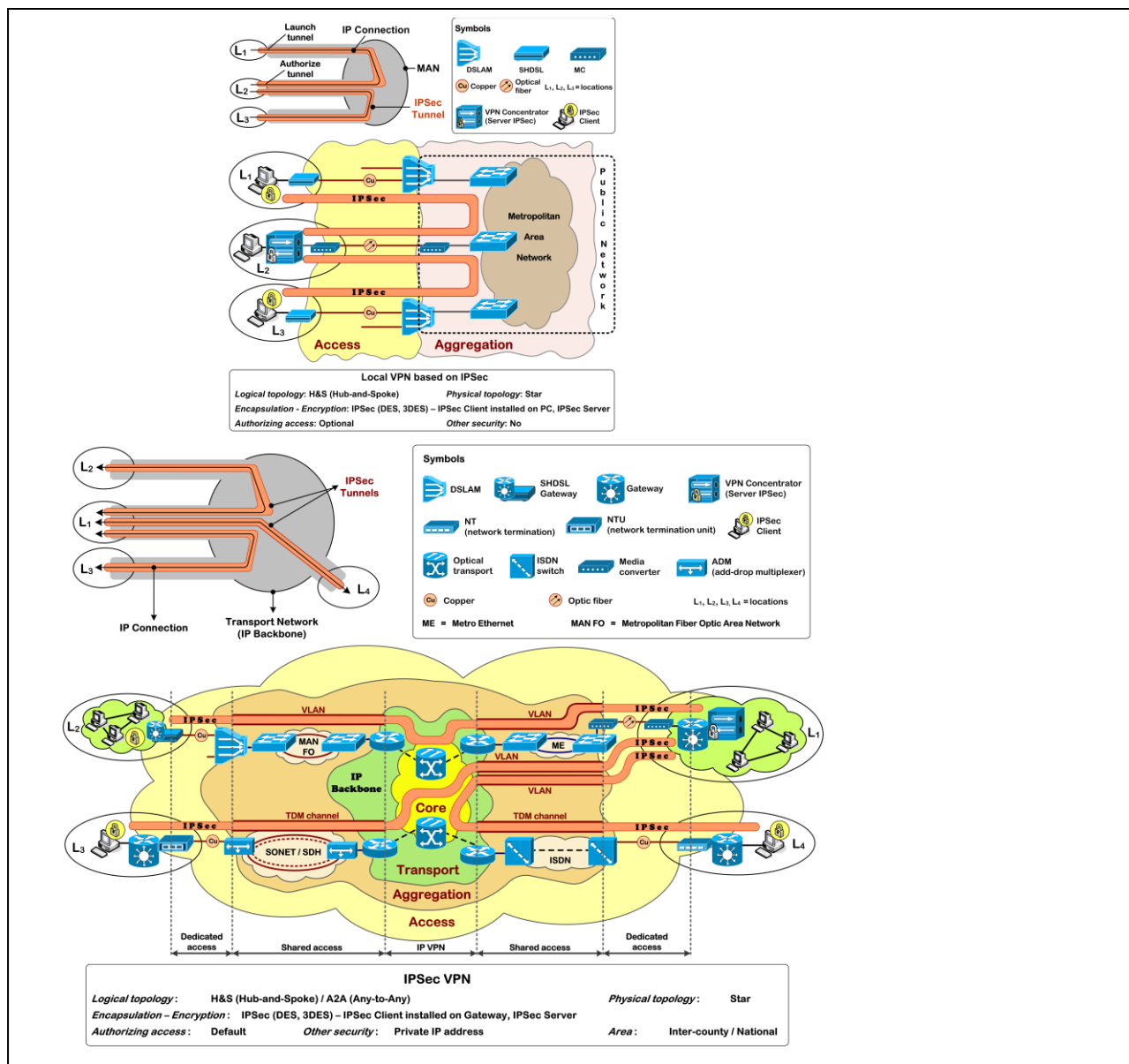


Figure 5. The local VPN & wide area VPN (e.g. national VPN) based on IPSec [15] [26]

IV. THE ENTERPRISE NETWORK MANAGEMENT

Enterprise networks are often large, run a wide variety of applications and protocols, and typically operate under strict reliability and security constraints; thus, they represent a challenging environment for network management [27]. Exactly network topology information is required to perform management activities (e.g. fault detection, root cause analysis, performance monitoring, load balancing, etc.) in enterprise networks. The importance of effective network management, not just in terms of controlling cost but in achieving the strategic aims of business, is also highlighted by some of the benefits respondents attributed to it, including improved inventory planning across the entire network, avoidance of *fire-fighting* situations by improved production and dispatch planning, reduced lead times and improved responsiveness and transparency at the enterprise level [28].

Network management represents the activities, methods, procedures, and tools (software and hardware) that pertain to the *operation, administration, maintenance, and provisioning* of networked systems. Method of solving this problem is to use a host and service monitor designed to inform as of network problems before your clients, end-users or managers do [15].

We implemented Nagios® to an enterprise part of an industrial group constituted as a dynamic alliance of many different small and medium sized enterprises (see figure 6 and 7) which has its headquarters in Bucharest and branch offices (agencies) in the country – in big cities but also in medium and small cities. All industrial holding locations have a local area network and communicate among themselves through a virtual private network. In each location were made two or three loops – one copper, one optical fiber and/or radio. The solution proposed and implemented by us was to use.

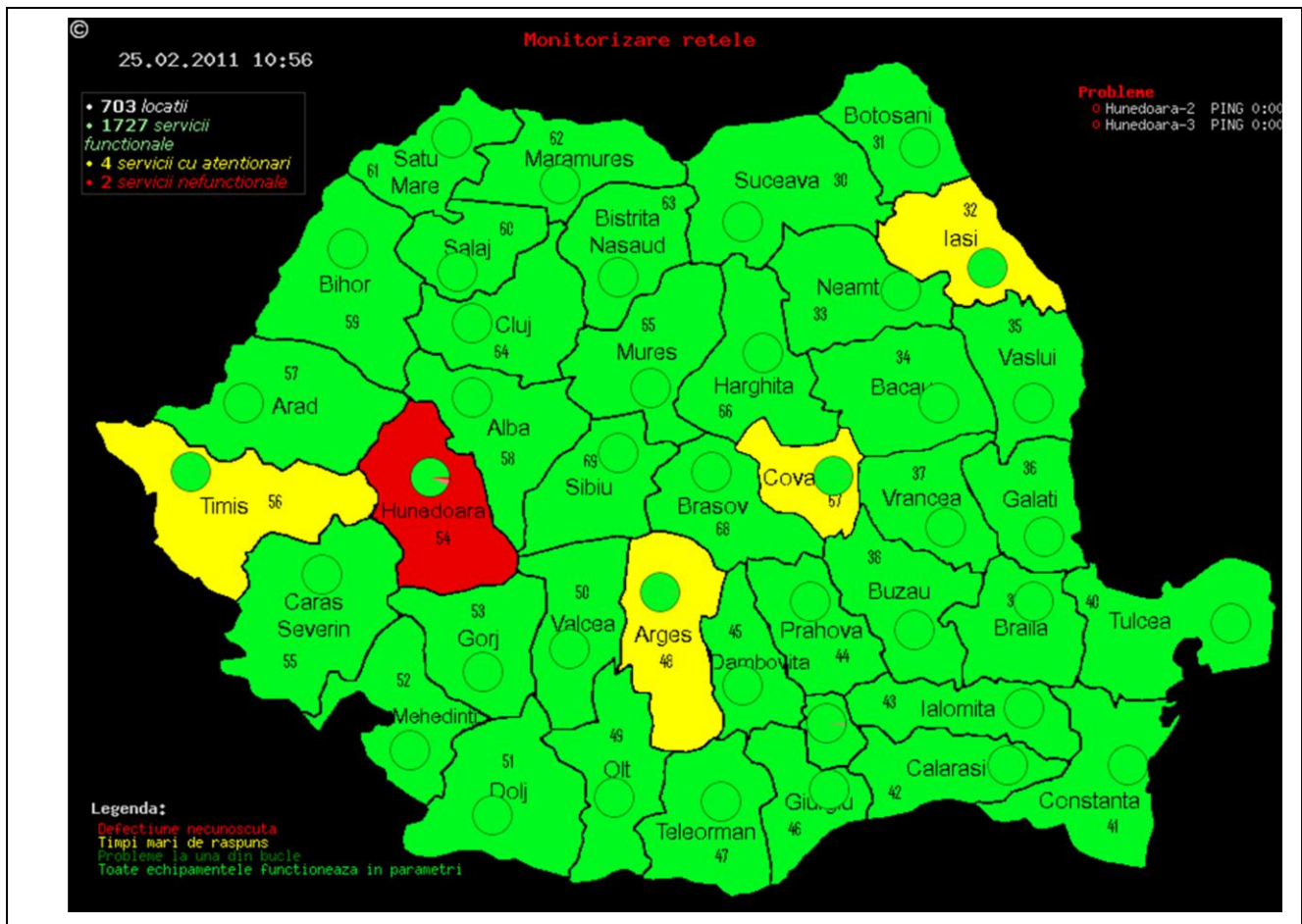


Figure 6. Status for an enterprise network – the local enterprise agency of Hunedoara County is down and the local agencies of Arges, Covasna, Iasi and Timis Counties have long response times

Nagios is a host and service monitor designed to inform administrators of network problems before your clients, end-users or managers do. It is based on queries that can be done at scheduled intervals with small programs called plug-ins. Those programs make queries on public services and return results as follows: 0 – OK, 1 – Warning and 2 – Critical. Nagios® architecture consists of a nucleus that collects data generated by plug-ins and notices on them, a number of plug-ins and an optional web interface where they are viewed, the state services and hosts monitored, history, notices sent, etc. In terms of performance, Nagios® scales very well and depending on the hardware configuration can verify tens of thousands of services. It can also be installed in cluster configuration. This software is licensed under the terms of the GNU General Public License Version 2 as published by the Free Software Foundation (GNU General Public License is a free, copy left license for software and other kinds of works). This gives legal permission to copy, distribute and/or modify Nagios® under certain conditions. Some of the many Nagios® features include [29]:

- Monitoring of network services (SMTP, POP3, HTTP, NNTP, ICMP) and monitoring of host resources (processor load, disk and memory usage, running

processes, log files, etc.) and monitoring of environmental factors such as temperature;

- Simple plug-in design that allows users to easily develop their own host and service checks and ability to define network host hierarchy, allowing detection of and distinction between hosts that are down and those that is unreachable;
- Contact notifications when service or host problems occur and get resolved (via email or other user-defined method);
- A Web interface – viewing current network status, notification and problem history, etc.

To implement this application we have used over five hundred locations. In figure 6 and 7 are presented the locations monitored for this large enterprise – we eliminate the beneficiary name for advertising reason. We've installed Nagios® in the PREMINV e-platform on a server with the following technical characteristics: 2 Dual Core Intel Xeon (TM) 3.6 GHz processors (64-bit), 2 GB RAM, 2 x 80 GB Hard Drives and Debian Linux 4.0 Operating System. We realized in the PREMINV e-platform more scripts as support for different operations.

Bihor_Marghita

5.59 0.18 : rt min/avg/max/mdev = 12.784/12.864/12.953/0.153 ms
Cisco IOS Software, 2801 Software (C2801-ADVIPSERVICESK9-M), Version 12.4(3b), RELEASE SOFTWARE (fc3)

Verificare		Rezultat	
Bucda	PING OK - Packet loss = 0%, RTA = 18.40 ms [rta=18.39800ms;1000.000000;2000.000000;0.000000 pl=0%;50;80;0		
Bucda	PING OK - Packet loss = 0%, RTA = 12.84 ms [rta=12.839000ms;1000.000000;2000.000000;0.000000 pl=0%;50;80;0		

Interfata	Administrativ	Operational
FastEthernet0/0	up(1)	up(1)
FastEthernet0/1	up(1)	up(1)
FastEthernet0/3/0	up(1)	up(1)
FastEthernet0/3/1	up(1)	down(2)
FastEthernet0/3/2	up(1)	down(2)
FastEthernet0/3/3	up(1)	down(2)
Null0	up(1)	up(1)
Vlan1	up(1)	down(2)
Tunnel1	up(1)	up(1)
Tunnel2	up(1)	up(1)
Tunnel100	up(1)	up(1)
Tunnel200	up(1)	up(1)
Vlan10	up(1)	up(1)
Async1	up(1)	down(2)

Figure 7. Status for an enterprise agency – the local enterprise agency of Bihor County providers loop status

V. CONCLUSIONS

In the actual context of the virtual enterprise network expanding, companies are much more preoccupied to build such structures and/or to be part of different structures that already exist. These will give them more business opportunities and by the knowledge transfer processes, they will gain competitiveness. Therefore, enterprises continue to implement information and communication technology systems solutions and strategies to improve their business processes in virtual networks.

Considering future product development as collaboration and communication oriented we implemented in the PREMINV e-platform a solution based on a virtual enterprise network (VPN IPSec solution) concept using integrated data sets and tools. As a general requirement for this virtual network based application the companies must be able to inter-operate and exchange data, information and knowledge in real time so that they can work as a single integrated unit, although keeping their autonomy. Also, virtual networked business teams need a strategic framework in which to operate. Today's virtual business teams don't appear to be able to fully leverage the much-touted opportunities offered by always-on interconnectedness, easy access to unlimited information sources and real-time communication tools. They also need good planning and in-depth project analysis, effective and accessible technologies, constant coaching, systematic fine-tuning, feedback processes and the full understanding that their success cannot be determined by a pre-designated set of communication technologies [26].

A solution for a large enterprise geographically dispersed network monitoring using open source software (Nagios®) has been presented in this paper. For an enterprise, network monitoring is a critical and very important function, which can save significant resources, increase network performance, employee productivity and maintenance cost of infrastructure. Nagios® compares the features and performance with expensive commercial monitoring applications as HP Operations Manager or Microsoft System Operations Manager. This software (Nagios®) can be developed and implemented at a corporate level but also in a company that provides telecommunication services.

This work was realized at the UPB-PREMINV Research Centre. The validation of this solution by a case study in the ORGVIRT & PROGPROC & ID 1022 research projects is to determine the new organization type for integrating the virtual enterprise medium and to outsource shared resources from UPB-PREMINV research centre to industrial partners.

REFERENCES

- [1] R. Ekwall and A. Schiper, "Replication: Understanding the Advantage of Atomic Broadcast over Quorum Systems", Journal of Universal Computer Science, vol. 11(2), pp. 703–711, 2005.
- [2] J. Niemann, S. Tickewitch and E. WestKamper, Design and Sustainable Product Life Cycles, Springer-Verlag Berlin Heidelberg, Germany, 2009.
- [3] J. B. Ayers, Handbook of supply chain management (2nd Ed.), Taylor & Francis Group, New York, USA, 2006.
- [4] R. Ferdiana and O. Hoseanto, "Instruction Design Model for Self-Placed ICT System E-Learning in an Organization", International Journal of Advanced Computer Science and Applications, vol. 3(8), pp. 1-7, 2012.

In figure 8 we present a script that save configurations for some Cisco® equipments. With this script one can connect the device (having entered at the command line user login and password) and copy line by line equipment configuration.

```

get_conf.php :
<?php
function get_conf($ip,$uname,$upass,$pass,$gr)
{
    $ip=$ip;

    $date=date('Ymd_His');
    $fname=$ip.'_'.$date;
    $fstart=fopen($gr.'/start_'.$fname.'.cfg','w');
    $frun=fopen($gr.'/run_'.$fname.'.cfg','w');
    $slink=fsckopen($ip,23,$errn,$errm);
    if(!$slink){
        // echo "Connection error"; exit;
        return 0;
    }
    else{
        // echo "IP : ".$ip;
        fputs($slink,$uname."\n");
        fputs($slink,$upass."\n");
        if(strlen($pass)){
            fputs($slink,"en\n");
            fputs($slink,$pass."\n");
        }
        fputs($slink,"\n\n\n");
        $gata=0;
        while($gata<1) {
            $ras=fgets($slink,128);
            // echo $ras."\n";
            if(strpos(" ".$ras,"invalid")>1){echo "\nUser/password wrong !\n";return 1;}
            if(strpos(" ".$ras,"Bad")>1){echo "\npassword enable wrong !\n";return 1;}
            if(strpos(" ".$ras,"#")>1){$gata=2;}
        }
        fputs($slink,"sh run\n");
        fputs($slink,"");
        $gata=0;$srun=0;
        while($gata<10000) {
            $ras=trim(fgets($slink,128));
            if(strncmp($ras,'A',1)<0){$ras=trim(substr($ras,8));}
            while((strlen($ras)>0) && (strncmp($ras,'A',1)<0)){$ras=substr($ras,1);}
        }
        // echo $ras."\n";
        if($srun){fputs($fstart,trim($ras)."\n");}
        fputs($slink,"");
        if(strlen($pass)>0){$gata=100000;}
        if(strlen($pass,'bytes')>0){$srun=1;}
        $gata++;
        fputs($slink,"\n\nsh start\n");
        fputs($slink,"");
        $gata=0;$sstart=0;
        while($gata<10000) {
            $ras=trim(fgets($slink,128));
            if(strncmp($ras,'A',1)<0){$ras=trim(substr($ras,8));}
            while((strlen($ras)>0) && (strncmp($ras,'A',1)<0)){$ras=substr($ras,1);}
        }
        // echo trim($ras)."\n";
        if($sstart){fputs($fstart,trim($ras)."\n");}
        fputs($slink,"");
        if(strlen($pass,>0) && (strlen($pass,'bytes')>0) && (strlen($pass,'bytes')>0){$sstart=1;}
        $gata++;
        fputs($slink,"\n\nnext\n");
        fputs($slink,"\n\nnext\n");
        fputs($slink,"\n\nnext\n");
        echo "$date OK!\n";
        fclose($fstart);fclose($frun);
        return $date;
    }
}
    
```

Figure 8. The Nagios® script example

- [5] T. Huang, G. Q. Wu and J. Wei, "Runtime monitoring composite Web services through state full aspect extension", *Journal of Computer Science and Technology*, vol. 24(2), pp. 294-308, 2009.
- [6] A. D. Potorac, "Considerations on VoIP Throughput in 802.11 Networks", *Advances in Electrical and Computer Engineering*, vol. 9(3), pp. 45-50, 2009.
- [7] C. H. Hauser, D. E. Bakken, I. Dionysiou, H. K. Gjermundrod, V. S. Irava, J. Helkey and A. Bose, "Security, trust and QoS in next-generation control and communication for large power system", *Int. J. Critical Infrastructures*, vol. 4(½), pp. 3-16, 2008.
- [8] Z. Hulicki, "Drivers and Barriers for Development of Broadband Access – CE Perspective", *Journal of Universal Computer Science*, vol. 14(5), pp. 717-730, 2008.
- [9] P. Sasvari, "The macroeconomic effect of the information and communication technology in Hungary", *International Journal of Advanced Computer Science and Applications*, vol. 2(12), pp. 75-81, 2011.
- [10] G. Dragoi, A. Draghici, S. M. Rosu, A. Radovici and C. E. Cotet, "Professional Risk Assessment Using Virtual Enterprise Network Support for Knowledge Bases Development", *Communications in Computer and Information Science*, vol. 110, part II, J.E. Quintela Varajao et al. (Eds.), Springer-Verlag Berlin Heidelberg, Germany, pp. 168-177, 2010.
- [11] L. J. Osterweil, "Formalism to support the definition of processes", *Journal of Computer Science and Technology*, vol. 24(2), pp. 198-211, 2009.
- [12] G. Dragoi, A. Draghici, S.M. Rosu and C. E. Cotet, "Virtual Product Development in University-Enterprise Partnership", *Information Resources Management Journal*, vol. 23(3), pp. 43-59, 2010.
- [13] A. Shakya, H. Takeda and V. Wuwongse, "StYLiD: Social information sharing with free creation of structured linked data", in *SWKM'2008: Workshop on Social Web and Knowledge Management @ WWW 2008*, April, Beijing, China, 2008.
- [14] M. Cudanov, O. Jasko and M. Jevtic, "Influence of Information and Communication Technologies on Decentralization of Organizational Structure", *Computer Science and Information System*, vol. 6(1), pp. 93-109, 2009.
- [15] S. M. Rosu and G. Dragoi, "VPN Solutions and Network Monitoring to Support Virtual Teams Work in Virtual Enterprises", *Computer Science and Information System*, vol. 8(1), pp. 1-26, 2011.
- [16] Cisco System, *Enterprise QoS Solution Reference Network Design Guide, Version 3.3*, Cisco Systems Inc., San Jose, CA, USA, 2008.
- [17] R. Skoog, A. Von Lehmen, G. Clapp, J. W. Gannett, H. Kobrinski and V. Poudyal, "Metro network design methodologies that build a next-generation network infrastructure based on this generation's services and demands", *Journal of Lightwave Technology*, vol. 22(11), pp. 2680-2692, 2004.
- [18] H. Youssef, S. M. Sait and S. A. Khan, "Topology design of switched enterprise networks using a fuzzy simulated evolution algorithm, *Engineering Applications of Artificial Intelligence*", vol. 15, pp. 327-340, 2002.
- [19] S. M. Rosu and G. Dragoi, "Virtual Enterprise Network General Architecture", in *Proceedings of the 8th International Conference on Communications*, pp. 313-316, Bucharest, Romania, ©IEEE, 2010.
- [20] A. G. Mason, *Cisco Secure Virtual Private Networks*, Published by Pearson Education, Cisco Press, 2001.
- [21] A. Moreno and K. Reddy, *Network virtualization*, Published by Pearson Education, Cisco Press, 2006.
- [22] S. M. Rosu, G. Dragoi, L. Rosu and M. Guran, "Virtual Enterprise Network Solutions to Support E-learning Sites Development", in *DAAAM International Scientific Book 2010*, chapter 63, B. Katalinic (Ed.), DAAAM International Press, Vienna, Austria, pp. 725-742, 2010.
- [23] J. Ward and J. Peppard, *Strategic planning for information systems*, John Wiley & Sons Press, West Sussex, UK, 2002.
- [24] A. Jamali, N. Naja and D. El Oudghiri, "An Enhanced MPLS-TE for Transferring Multimedia packets", *International Journal of Advanced Computer Science and Applications*, vol. 3(8), pp. 8-13, 2012.
- [25] R. Deal, *The Complete Cisco VPN Configuration Guide*, Published by Pearson Education, Cisco Press, 2005.
- [26] S. M. Rosu and G. Dragoi, "Virtual Enterprise Network Solutions and Monitoring as Support for Geographically Dispersed Business", in *Handbook of Research on Business Social Networking: Organizational, Managerial and Technological Dimensions*, IGI Global, Hershey, PA, USA, pp. 34-62, 2012.
- [27] M. Casado, M. J. Freeman, J. Pettit, J. Luo, N. Gude, N. McKeown and S. Shenker, "Rethinking enterprise network control", *Transactions on Networking*, vol.17, no. 4, pp. 1270-1283, 2009.
- [28] N. Jones and Q. Wang, "A mixed integer programming approach for logistic network design and optimization information and value adding networks", in *DET2009 Proceedings, AISC66*, Huang G et al. (Eds.), Springer-Verlag Berlin Heidelberg, pp. 1227-1241, 2010.
- [29] E. Galstad, *Nagios® Version 2.x Documentation*, Published by www.nagios.org, 1999-2006.

AUTHORS PROFILE

Dr. Sebastian Marius Rosu received the B.E. in Aerospace Construction and in Informatics from University "Polytechnica" of Bucharest, in 2000 and University „Dunărea de Jos” of Galați in 2002, respectively. In 2009 obtained doctoral degree in Automatics from University "Polytechnica" of Bucharest. He is telecommunications engineer at the Romanian Special Telecommunications Service. He has been active in the fields of the quality control, risk management, industrial informatics, knowledge management, computers, computer integrated enterprise, computer networks, networks design and collaborative design systems. Since 2004, he has been associated with the PREMINV Research Laboratory at the Polytechnic University of Bucharest first as PhD Student and from 2009 as PhD.

Marius Marian Popescu received the B.E. in Electronics from Military Technical Academy of Bucharest. Currently he is PhD candidate at University of Pitești, Faculty of Electronics, Communications and Computers. Since as 2001 he is IT engineer at the Romanian Special Telecommunication Service. His interest areas are pattern recognition, network security, network operations center development, network project management and network equipment (VPN, IPsec).

Dr. George Dragoi received the B.S. in Electronics & Telecommunications and doctoral degrees in Industrial Informatics from University "Polytechnica" of Bucharest, in 1982 and 1994, respectively. From 1982-1986, he was an Associate Researcher of the Romanian National Research Center in Telecommunication of Bucharest. From 1986 to 1990 he was a Research Assistant, from 1990 to 1993 lecturer, from 1993 to 1998 associate professor and from 2006 to present professor at the "Politehnica" University of Bucharest, Faculty of Engineering in Foreign Languages. He has been active in the fields of the industrial robots control, industrial informatics, computer integrated enterprise, computer networks and telecommunications, virtual enterprise and collaborative design systems. Since 1998, he has been associated with the Institute National Polytechnic at Grenoble, France, as Associate and Visiting Professor. He received the Romanian Academy Award "Aurel Vlaicu" in 1993 and it has served as a member of the National Council for Technological Education and Innovation after 1999.

Ioana Raluca Guica is PhD candidate at University "Polytechnica" of Bucharest. Currently she is assistant professor at the "Politehnica" University of Bucharest, Faculty of Engineering in Foreign Languages.

Analyzing the Efficiency of Text-to-Image Encryption Algorithm

Ahmad Abusukhon
Computer Network Department
Al-Zaytoonah University of Jordan
Amman, Jordan

Mohammad Talib
Department of Computer Science
University of Botswana
Gaborone, BOTSWANA

Maher A. Nabulsi
Department of Computer Science
Al-Zaytoonah University of Jordan
Amman, Jordan

Abstract—Today many of the activities are performed online through the Internet. One of the methods used to protect the data while sending it through the Internet is cryptography. In a previous work we proposed the Text-to-Image Encryption algorithm (TTIE) as a novel algorithm for network security. In this paper we investigate the efficiency of (TTIE) for large scale collection.

Keywords-Algorithm; Network; Secured Communication; Encryption & Decryption; Private key; Encoding.

I. INTRODUCTION

Cryptography or sometimes referred to as encipherment is used to convert the plaintext to encode or make unreadable form of text [1]. The sensitive data are encrypted on the sender side in order to have them hidden and protected from unauthorized access and then sent via the network. When the data are received they are decrypted depending on an algorithm and zero or more encryption keys as described in Fig.1.

Decryption is the process of converting data from encrypted format back to their original format [2]. Data encryption becomes an important issue when sensitive data are to be sent through a network where unauthorized users may attack the network. These attacks include IP spoofing in which intruders create packets with false IP addresses and exploit applications

that use authentication based on IP and packet sniffing in which hackers read transmitted information.

One of the techniques that are used to verify the user identity (i.e. to verify that a user sending a message is the one who claims to be) is the digital signature [3]. Digital signature is not the focus of this research.

There are some standard methods which are used with cryptography such as private-key (also known as symmetric, conventional, or secret key), public-key (also known as asymmetric), digital signature, and hash functions [4]. In private-key cryptography, a single key is used for both encryption and decryption. This requires that each individual must possess a copy of the key and the key must be passed over a secure channel to the other individual [5]. Private-key algorithms are very fast and easily implemented in hardware; therefore, they are commonly used for bulk data encryption.

The main components of the symmetric encryption include - plaintext, encryption algorithm, secret key, ciphertext and decryption algorithm. The plaintext is the text before applying the encryption algorithm. It is one of the inputs to the encryption algorithm. The encryption algorithm is the algorithm used to transfer the data from plaintext to ciphertext. The secret key is a value independent of the encryption

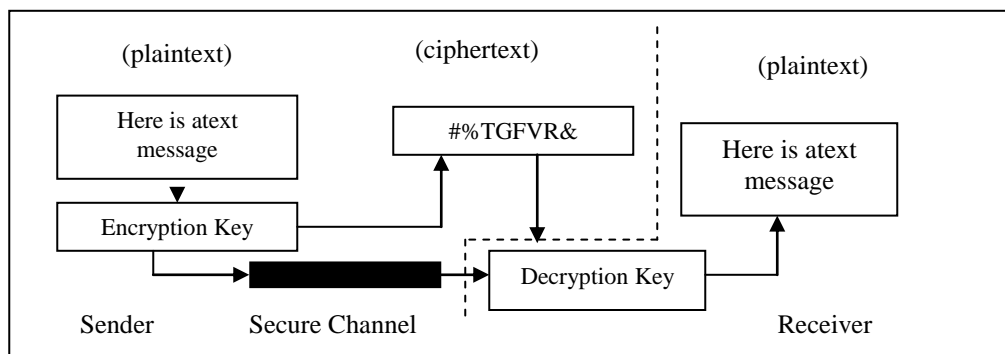


Figure 3. Encryption and Decryption methods with a secure channel for key exchange.

algorithm and of the plaintext and it is one of the inputs of the encryption algorithm. The ciphertext is the scrambled text produced as output. The decryption algorithm is the encryption algorithm run in reverse [6, 2, 7]. There are two main categories of private-key algorithms, namely block and stream encryption [8].

Public-key encryption uses two distinct but mathematically related keys - public and private. The public key is the non-secret key that is available to anyone you choose (it is often made available through a digital certificate). The private key is kept in a secure location used only by the user. When data are sent they are protected with a secret-key encryption that was encrypted with the public key. The encrypted secret key is then transmitted to the recipient along with the encrypted data. The recipient will then use the private key to decrypt the secret key. The secret key will then be used to decrypt the message itself. This way the data can be sent over insecure communication channels [6].

II. RELATED WORK

Bh, Chandravathi, and Roja [9] proposed encoding and decoding a message in the implementation of Elliptic Curve Cryptography is a public key cryptography using Koblitz's method [10, 11]. In their work, each character in a message is encoded by its ASCII code then the ASCII value is encoded to a point on the curve. Each point is encrypted to two ciphertext points. Our work differs from their work. In their work they used public-key technique whereas in our work we use private key technique. They encoded each character by its ASCII value but we encode each character by one pixel (three integer values - R for Red, G for Green and B for Blue).

Singh and Gilhorta [5] proposed encrypting a word of text to a floating point number that lie in range 0 to 1. The floating point number is then converted into binary number and after that one time key is used to encrypt this binary number. In this paper, we encode each character by one pixel (three integer values R, G and B).

Kiran Kumar, MukthiarAzam, and Rasool [12] proposed a new technique of data encryption. Their technique is based on matrix disordering which was relied on generating random numbers used for rows or columns transformations. In their work, the original plaintext was ordered into a Two-directional circular queue in a matrix A of order $m \times n$. A number of column and row transformations were carried-out on the matrix and to do so a random function was used to generate positive integer say X and then X is converted to a binary number. Rows or columns transformation was made based on the values of the individual bits in the binary number resulted from the X value.

Another random number was generated in order to determine the transformation operation. The random number was divided by three (as we have three types of transformation operations) and the modulus (0, 1, or 2) was used to determine the operation type. The operation type could be 0 (means circular left shift), 1 (means circular right shift) and 2 (means reverse operation on the selected rows). In case rows were selected to perform a transformation operation (the selection was made depending on the bit value of X) two random numbers r1 and r2 were generated where r1 and r2 represent two distinct rows. Another two random numbers were generated c1 and c2 that represent two distinct columns. The two columns c1 and c2 were generated in order to determine the range of rows in which transformation had to be performed. After the completion of each transformation a sub-key was generated and stored in a file key which was sent later to the receiver to be used as decryption key. The sub-key format is T, Op, R1, R2, Min, and Max, where:

T: the transformation applied to either row or column

Op: the operation type coded as 0, 1, or 2, e.g., shift left array contents, shift right array contents, and reverse array contents.

R1 and R2: two random rows or columns

Min, Max: minimum and maximum values of a range for two selected rows namely, R1 and R2.

Abusukhon and Talib [13] proposed a novel data encryption algorithm called Text-to-Image Encryption algorithm (TTIE) in which a given text is encrypted into an image. Each letter from the plaintext is encrypted into one pixel. Each pixel consists of three integers and each integer represents one color (for example, Red, Green, and Blue). Each color has a value in the range from 0 to 255. The private key is generated randomly by creating three random integers (one pixel) for each letter. For example, letter "A" may be represented by the RGB value (0, 7, 0). The whole letters of a given text is transformed into a two dimensional array of pixels say M, then M is shuffled a number of times by performing row and column swapping. In their work, they analyzed the TTIE algorithm by calculating the number of possible permutations to be guessed by hackers. The TTIE algorithm is described in Fig. 2. In their work, they mentioned that the TTIE algorithm is useful for text encryption for individual offline machines, a network system, and for e-mail security. They proposed the TTIE algorithm for e-mail security since the text messages in the mail box appear as images making it difficult for others to guess the plaintext messages (i.e. the original messages sent from the other side).

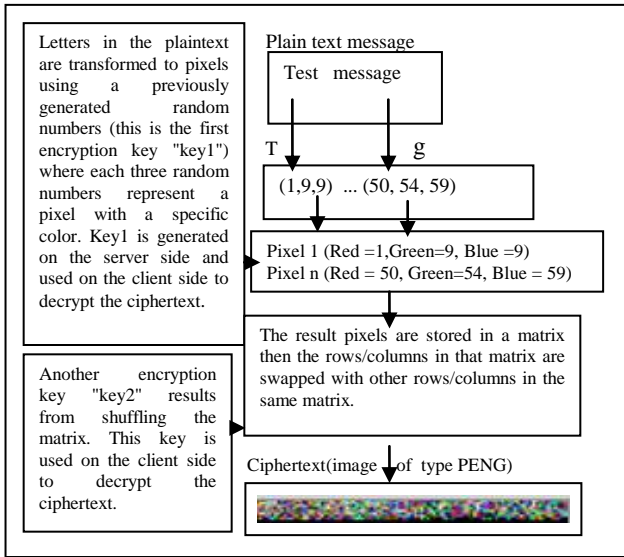


Figure 2. The Text-to-Image Encryption Algorithm

In this paper, we propose to investigate the efficiency of the TTIE algorithm when various memory sizes and various data sizes are used.

III. OUR EXPERIMENTS

Abusukhon and Talib [13] tested the TTIE algorithm for a few documents of size 512 KB. In this paper, we propose to investigate the efficiency of the TTIE algorithm when different data sizes are used while fixing the memory size to 250MB. In addition, we propose to investigate the effect of using various memory sizes on the performance of the TTIE algorithm while fixing the data collection size to 1.96 Gigabytes. We use Java NetBeans as a vehicle to carry out our experiments.

A. Machine Specifications

Our experiments are carried out on a single machine with the following specifications; processor Intel (R) core (TM)2, Duo CPU T5870 @ 2.00GHz, installed memory (RAM) 2.00GB operating system Windows 7 professional and hard disk 25.2 GB (free space).

B. Data Collection

We build the data collection for our experiments by writing a simple Java code. This code is used to generate documents of different sizes. This is done by setting the length of the word and the number of words in each document. The letters in each document are chosen randomly from an array that contains all letters from "A" to "Z". For our experiments, we set the word length to 7 and the number of words in each document to 30.

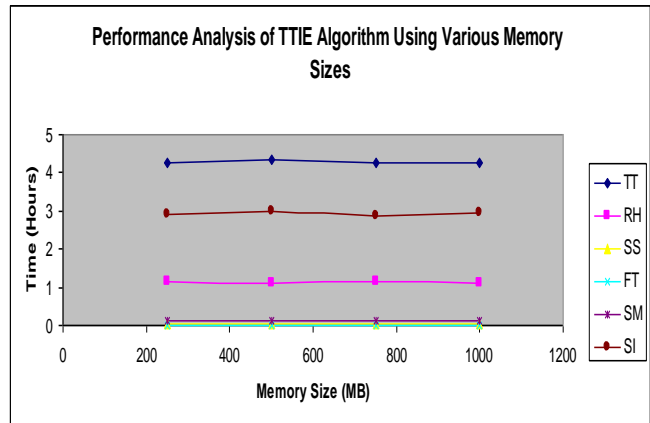
C. Investigate the Effect of Memory Size on Performance

In this section, we investigate how the TTIE algorithm is affected by the memory size. To do so, we divide the time of the TTIE algorithm into five basic times namely, the time required for reading the data collection from the hard disk (RH), the time of the switch statement -switch statement is used to transfer the letters in a given text into pixels and store

them into a single dimension array- (SS), the time required to fill the array of pixels into a two dimension array – filling pixels in a two dimensional array is necessary to perform matrix scrambling- (FT), the time required for scrambling the matrix (SM), the time required to store the result image (the encrypted text) on the hard disk (SI). We set the collection size in all experiments carried out in this section to 1.96 GB. We calculate the total time (TT) of each experiment as well as the five basic times mentioned above (i.e. RH, SS, FT, SM, and SI). We investigate how the TTIE is affected by various memory sizes 250MB, 500MB, 750MB, and 1000MB as described in Fig. 3.

Fig. 3, shows that the most dominant time of the TTIE algorithm (with respect to the TT time) is the SI time. The average of the total time (TT) for all experiments carried out in this section is 4.289. However, the average of the SI time for all experiments carried out in this section is 2.922. Thus the SI time is 0.681 of the TT time. In addition, Fig. 3 shows that the TT time and the SI time are slightly affected by changing the memory size.

Figure 3. The performance of the TTIE algorithm using various memory



sizes

D. Investigate the Effect of Data Size on Performance

In this section, we investigate how the performance of the TTIE algorithm is affected by various data sizes while fixing the memory size to 250MB. We use different data collection sizes; 118MB, 236MB, 354MB, 473MB, 591MB, 709MB, 827MB, 946MB, 1.03GB, and 1.15GB. Fig. 4 describes the results of these experiments.

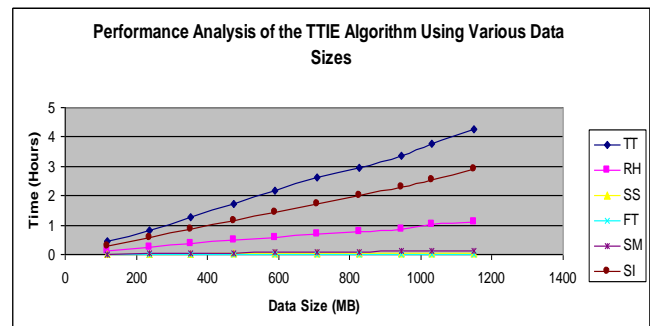


Figure 4. The performance of the TTIE algorithm using various data sizes

Our results show that the most dominant time of the TTIE algorithm is the SI time. The average of the total time (TT) for all experiments carried out in this section is 2.337. However, the average of the SI time for all experiments carried out in this section is 1.577. This means that most of the TTIE time is spent on saving the images on the hard disk.

IV. CONCLUSION AND FUTURE WORK

In this paper, we investigated the efficiency of the TTIE algorithm. The TTIE algorithm is good for text encryption for a network system (TTIE is good for a Virtual Private Network, VPN, where encrypted data are sent and received across shared or public networks), individual offline machines, and e-mail security. We investigated the efficiency of the TTIE algorithm when various memory sizes and various data sizes are used. The results from our experiments showed that the most dominant time is the time required to save the encrypted data (images) on the hard disk (this time is called SI). In addition, the results from our experiments showed that the SI time, SM time, FT time, SS time, RH time, and the TT time are very slightly changed when the memory size is increased. In addition, the results from this work showed that the SI time, TT time, and the RH time are greatly increased when the size of the data collection is increased. The SS time, FT time, and the SM time are slightly changed when the size of the data collection is increased. In future, we propose to reduce the time required for saving the encrypted data (images) on the hard disk using distributed computing.

We propose to use a large data collection size (multiple Gigabytes) and investigate distributing the data collection among a number of nodes working together with a server. All nodes work in parallel to encrypt and store the large data collection on the hard disk.

ACKNOWLEDGMENT

Heartfelt gratitude to Al-Zaytoonah Private University of Jordan for their help and for their financial support to carry out this work successfully.

REFERENCES

- [1] K. Lakhtaria, "Protecting computer network with encryption technique: A Study". International Journal of u- and e-service, Science and Technology, vol. 4, 2011.
- [2] A. Chan, "A Security framework for privacy-preserving data aggregation in wireless sensor networks". ACM transactions on sensor networks, vol.7, 2011.
- [3] M. Savari, and M. Montazerolzhour, "All About encryption in smart card". Proceeding of International conference on Cyber Security, Cyber Warfare and Digital Forensic (Cyber Sec), 2012.
- [4] B. Zaidan, A. Zaidan, A. Al-Frajat, and H. Jalab, "On the differences between hiding information and cryptography techniques: An Overview". Journal of Applied Sciences vol. 10, 2010.
- [5] A. Singh, and R. Gilhorta, "Data security using private key encryption system based on arithmetic coding". International Journal of Network Security and its Applications (IJNSA), vol. 3, 2011.
- [6] W. Stallings, Cryptography and network security principles and practices ,4th ed. Prentice Hall. [online] at: <http://www.filecrop.com/cryptography-and-network-security-4th-edition.html>, Accessed on 1-Oct-2011.
- [7] V. Gupta, G. Singh, and R. Gupta, "Advance cryptography algorithm for improving data security". vol. 2, 2012. [online] at: <http://www.ijarcsse.com/docs/papers/january2012/V2I11055.pdf>, Accessed on 19-Nov-2012.
- [8] S. Suliman, Z. Muda, and J. Juremi, "The New approach of Rijndael Key schedule". Proceeding of International conference on Cyber Security, Cyber Warfare and Digital Forensic (Cyber Sec), 2012.
- [9] P. Bh, D. Chandravathi, and P. Roja, "Encoding and decoding of a message in the implementation of Elliptic Curve cryptography using Koblitz's method". International Journal of Computer Science and Engineering, vol.2,2010.
- [10] N. Koblitz, "Elliptic Curve cryptosystems", Mathematics of Computation", vol.48,1987,pp.203-209.
- [11] N. Koblitz, A Course in Number Theory and cryptography. 2nd ed. Springer-Verlag, 1994.
- [12] M. Kiran Kumar, S. Mukthiar Azam, and S. Rasool, "Efficient digital encryption algorithm based on matrix scrambling technique". International Journal of Network Security and its Applications (IJNSA), vol.2,2010.
- [13] A. Abusukhon, and M. Talib, "A Novel network security algorithm based on Private Key Encryption". In Proceeding of The International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec12), 2012.

DNA Sequence Representation and Comparison Based on Quaternion Number System

Hsuan-T. Chang, Chung J. Kuo
Photonics and Information Lab,
Department of Electrical
Engineering, YunTech
Douliu Yunlin, 64002 Taiwan

Neng-Wen Lo
Department of Animal Science and
Biotechnology,
Tunghai University
Taichung, 40704 Taiwan

Wei-Z.Lv
Computer and Communication Lab,
Industrial Technology Research
Institute
Hsinchu, 310 Taiwan

Abstract—Conventional schemes for DNA sequence representation, storage, and processing are usually developed based on the character-based formats. We propose the quaternion number system for numerical representation and further processing on DNA sequences. In the proposed method, the quaternion cross-correlation operation can be used to obtain both the global and local matching/mismatching information between two DNA sequences from the depicted one-dimensional curve and two-dimensional pattern, respectively. Simulation results on various DNA sequences and the comparison result with the well-known BLAST method are obtained to verify the effectiveness of the proposed method.

Keywords- *Bioinformatics; genomic signal processing; DNA sequence; quaternion number; data visualization.*

I. INTRODUCTION

Recently, the great progress on biotechnology makes the deoxyribonucleic acid (DNA) sequencing more efficiently. Huge amount of DNA sequences of various organisms have been successfully sequenced with higher accuracy. Analyzing DNA sequences can investigate the biological relationships such as homologous and phylogeny of different species. However, the analysis of DNA sequences using the biological methods is too slow for processing huge amount of DNA sequences. Therefore, the assistance of computers is necessary and thus *bioinformatics* is extensively developed. Efficient algorithms are desired to deal with the considerable and tedious biomolecular data.

Computer-based algorithms have solved various problems dealt in bioinformatics, such as the sequence matching (two and multiple sequences, global and local alignments), fragments assembly of DNA pieces, and physical mapping of DNA sequences. Most of the algorithms consider the data structures of DNA sequences as the string, tree, and graph. The artificial intelligence techniques such as the genetic algorithm [1], artificial neural networks [2], and data mining [3] have been intensively employed in this research area. In [4], the study of genomic signals mainly at scales of 104~108 bp, to detect general trends of the genomic signals, potentially significant in revealing their basic properties and to search for specific genomic signals with possible control functions. On the other hand, many distributed databases over the Internet have been constructed and can be easily accessed from the World Wide Web [5]-[7]. Most of the techniques treat the DNA sequences as the symbolic data, which are the

composition of four characters A, G, C, and T corresponding to the four types of nucleic acids: Adenine, Guanine, Cytosine, and Thymine, respectively. However, the biomolecular structures of genomic sequences can be represented as not only the symbolic data but also the numeric form.

Recently the genomic signal processing era has received a great deal of attention [8], [9]. The well-known digital signal processing (DSP) techniques have been developed to analyze the numeric signals for many applications [10]. If the symbolic DNA sequences can be transformed to the numeric ones, then the DSP-based algorithms would provide alternative solutions for the bioinformatics problems defined in the symbolic domain. Hsieh *et. al.* proposed DNA-based schemes to efficiently solve the graph isomorphism problems [11], [12]. Some previous studies have shown various methods of mapping the symbolic DNA sequences to numeric ones for further processing such as discrete Fourier transform or wavelet transform [13]-[15]. Then, the periodic patterns existed in DNA sequences can be observed from the determined scalograms or spectrograms. In assigning the four bases as some real or complex numbers such as $\pm 1 \pm i$, the further mathematical operations are straightforward and simple. However, exploring the biological relationship is difficult in the mapping between the bases and numbers.

Magarshak proposed a quaternion representation of RNA sequences [16]. Four bases with eight biological states form the group of quaternions and the tertiary structure of the RNA sequence can be analyzed through the quaternion formalism. Hypercomplex signals can be considered as the general form of quaternion signals [17]. Shuet *al.* proposed hypercomplex number representation for pairwise alignment and determining the cross-correlation of DNA sequences [18]-[20]. The DNA sequences are aligned with fuzzy composition and a new scoring system was proposed to adapt the hypercomplex number representation. Based on the similar ideas shown above, we adapt two implications using the quaternion number system [21] for DNA sequences. The quaternion numbers are complex numbers with one real part and three imaginary parts. Here four bases in a DNA sequence are assigned with four different quaternion numbers. A DNA sequence can thus be transformed into a quaternion-number sequence. Instead of finding the local frequency information of DNA sequences, the cross-correlation algorithm based on quaternion numbers is proposed for both the global and local matching between two DNA sequences. Since the real parts of four quaternion

numbers are all zero and the imaginary parts are with certain properties. The matching information can be observed from the real part in the result of the quaternion cross-correlation operation. In addition to the global cross-correlation result, the local matching information can be extracted from the product of each multiplication in the correlation operation. The global and local comparisons can be represented by 1-D curve and 2-D pattern, respectively. The simulation results show that the proposed quaternion number system can efficiently represent DNA sequences and then be used to determine the global and local sequence alignment with the help of the cross-correlation operation.

The organization of this paper is as follows: Section 2 introduces the basics of quaternion number systems and the quaternion number representation of DNA sequences. The global and local sequence matching based on the quaternion correlation is described. Section 3 provides the simulation results for certain DNA sequences, which verify the effectiveness of the proposed method. Finally, the conclusion is drawn in Section 4.

II. QUATERNION CORRELATION FOR SEQUENCE MATCHING

A. Quaternion Number Sequence Representation

Quaternion numbers [21] (also called the *hyper-complex numbers*) are the generalization of complex numbers. They have been applied in certain applications such as the color image filtering [22] and segmentation [23] and, the design of 3-D infinite impulse response filters [24]. Since the quaternion number system is not well known in all signal processing areas, here their properties are briefly reviewed.

The quaternion number has four components: one real part and three imaginary parts. The notation of a quaternion number q is defined as

$$q = q_a + \hat{i}q_b + \hat{j}q_c + \hat{k}q_d \quad (1)$$

Where q_a, q_b, q_c, q_d are real numbers, and $\hat{i}, \hat{j}, \hat{k}$ are operators for the three imaginary parts. The conjugate of quaternion number, q^* , is defined as

$$q^* = q_a - \hat{i}q_b - \hat{j}q_c - \hat{k}q_d \quad (2)$$

More detailed description of quaternion operations can be referred in [25].

Table I shows the proposed mapping method between the four characters and the corresponding quaternion numbers. Note that all the real parts are zero, while the imaginary parts can be considered as the coordinates of four vertices of a regular tetrahedron in 3-D space.

That is, the 3-D coordinates (x,y,z) of the vertices of a regular tetrahedron are applied to the imaginary part 'b,c,d' of quaternion numbers, and the real part 'a' of quaternion numbers are set to zero. Then, a character sequence is mapped to a quaternion number sequence. For example, a character sequence $s_c[n]=\{A,A,T,A,G,C,G,T\}$ is mapped to a quaternion number sequence $s_q[n]=\{q_A, q_A, q_T, q_A, q_G, q_C, q_G, q_T\}$.

TABLE I: THE MAPPING TABLE FROM A, C, T, AND G TO CORRESPONDING QUATERNION NUMBERS.

	q_a	q_b	q_c	q_d
A	0	0	0	1
T	0	$\frac{2\sqrt{2}}{3}$	0	$-\frac{1}{3}$
C	0	$-\frac{\sqrt{2}}{3}$	$\frac{\sqrt{6}}{3}$	$-\frac{1}{3}$
G	0	$-\frac{\sqrt{2}}{3}$	$-\frac{\sqrt{6}}{3}$	$-\frac{1}{3}$

After the mapping procedure in the former discussion, a quaternion number sequence standing for a DNA sequence is obtained:

$$s_q[n] = \{q_1, q_2, q_3, \dots, q_n\} \quad (3)$$

Then, an accumulating process is applied to obtain another quaternion number series

$$\bar{s}_q[n] = \{\bar{q}_1, \bar{q}_2, \bar{q}_3, \dots, \bar{q}_n\}, \quad (4)$$

where

$$\bar{q}_l = \sum_{n=1}^l q_n \quad (5)$$

By extracting the imaginary part of series $\bar{s}_q[n]$, the 3-D coordinates for each accumulated quaternion number in the series can be obtained. Line segments are used to connect these points in order, and a 3-D trajectory can thus be obtained for DNA sequence visualization [26]-[28].

B. Quaternion Correlation

Once the quaternion number sequences of two DNA sequences have been obtained, the cross-correlation operation is performed for the global comparison. The cross-correlation of two quaternion number sequences $s_{q_1}[n]$ and $s_{q_2}[n]$ of lengths M and N , respectively, is defined as

$$r_{1,2}[\xi] = \sum_{n=-\infty}^{\infty} s_{q_1}[n]s_{q_2}^*[n+\xi], \quad (6)$$

where ξ is the index of the correlation function and $0 \leq \xi \leq N+M-1$.

In the correlation operation, the conjugate operation is applied on one of the two quaternion numbers multiplied. Therefore, the cross-correlation of two identical and different symbols contributes +1 and -1/3 to the result, respectively. If there are two sequences to be correlated with length N and M ($N>M$), the real part of correlation result, v_{re} , for zbp overlap ($0 < z \leq M$) is

$$v_{re} = p + (z - p) \times (-1/3), \quad (7)$$

Where p is the matching counts and $z-p$ is the mismatching counts.

Equation (6) shows that the cross-correlation result for a specific ξ value is the sum of the product of the original and the shifted and conjugate sequences. The products in certain n are non-zero and zero values when two sequences overlap and not, respectively.

When the two sequences overlap, the product of two quaternion numbers reflects whether they are the same or not. That is, during the cross-correlation operation, the local alignment proceeds under a given ξ value. For example, if there are z overlapping numbers between two sequences, Eq. (6) becomes

$$r_{1,2}[\xi] = \sum_{n=m}^{m+z-1} s_{q_1}[n]s_{q_2}^*[n+\xi], \text{ for } z \geq 1, \quad (8)$$

Where m denotes the starting position of the overlap region under the given ξ value. Since there are $(N+M-1)$ possible ξ values in the cross-correlation result, all the possible local alignments between two sequences can be obtained. Therefore, the local alignment results can be shown in a 2-D array of size $(N+M-1) \times (N+M-1)$, in which the entry denotes the matching status of two nucleotides from two DNA sequences. A grayscale image $f_{Re}(x,y)$ of size $(N+M-1) \times (N+M-1)$ corresponding this 2-D array can be generated based on the following rule:

$$f_{Re}(x,y) = \begin{cases} 0, & \text{if } \text{Re}\{s_{q_1}[n]s_{q_2}^*[x+y]\} = 1, & (\text{overlap \& match}) \\ 255, & \text{if } \text{Re}\{s_{q_1}[n]s_{q_2}^*[x+y]\} = -\frac{1}{3}, & (\text{overlap but mismatch}) \\ 128, & \text{if } \text{Re}\{s_{q_1}[n]s_{q_2}^*[x+y]\} = 0, & (\text{non-overlap}) \end{cases} \quad (9)$$

Here $\text{Re}\{\cdot\}$ denotes the real part of the complex value in the bracket and $0 < x, y \leq N+M-1$. If the connected pixels in the horizontal direction constitutes as a black line in the image, the local matching between two sequences exists. Therefore, the local matching information between two DNA sequences can be obtained from the quaternion correlation result in addition to the global matching information.

TABLE II. THE IMAGINARY PARTS OF THE MULTIPLIED VALUES OF EVERY TWO DIFFERENT QUATERNION NUMBERS.

	AxT	AxC	AxG	TxC	TxG	CxG
\hat{i}	0	$-\frac{\sqrt{6}}{3}$	$\frac{\sqrt{6}}{3}$	$\frac{\sqrt{6}}{9}$	$-\frac{\sqrt{6}}{9}$	$-\frac{2\sqrt{6}}{9}$
\hat{j}	$\frac{2\sqrt{2}}{3}$	$-\frac{\sqrt{2}}{3}$	$-\frac{\sqrt{2}}{3}$	$\frac{\sqrt{2}}{3}$	$\frac{\sqrt{2}}{3}$	0
\hat{k}	0	0	0	$\frac{4\sqrt{3}}{9}$	$-\frac{4\sqrt{3}}{9}$	$\frac{4\sqrt{3}}{9}$

	TxA	CxA	GxA	CxT	GxT	GxC
\hat{i}	0	$\frac{\sqrt{6}}{3}$	$-\frac{\sqrt{6}}{3}$	$-\frac{\sqrt{6}}{9}$	$\frac{\sqrt{6}}{9}$	$\frac{2\sqrt{6}}{9}$
\hat{j}	$-\frac{2\sqrt{2}}{3}$	$\frac{\sqrt{2}}{3}$	$\frac{\sqrt{2}}{3}$	$-\frac{\sqrt{2}}{3}$	$-\frac{\sqrt{2}}{3}$	0
\hat{k}	0	0	0	$-\frac{4\sqrt{3}}{9}$	$\frac{4\sqrt{3}}{9}$	$-\frac{4\sqrt{3}}{9}$

C. Mismatching Analysis

The real part of correlation result reflects the matching information, while the imaginary part reflects the mismatching

information in sequence comparison. Therefore, the number of the matching and mismatching counts from the real and imaginary parts of the correlation result could be investigated. Let q_3 and q_4 denote the products of two quaternion numbers q_1 and q_2 . That is, $q_3 = q_1 \times q_2$ and $q_4 = q_2 \times q_1$, where $q_n = q_{a_n} + \hat{i}q_{b_n} + \hat{j}q_{c_n} + \hat{k}q_{d_n}$, $n=1,2,3,4$. According to the rules of quaternion multiplication, the following results can be obtained:

$$q_{a_3} = -q_{b_1} \times q_{b_2} - q_{c_1} \times q_{c_2} - q_{d_1} \times q_{d_2}, \quad (10)$$

$$q_{b_3} = q_{c_1} \times q_{d_2} - q_{d_1} \times q_{c_2}, \quad (11)$$

$$q_{c_3} = -q_{b_1} \times q_{d_2} + q_{d_1} \times q_{b_2}, \quad (12)$$

$$q_{d_3} = q_{b_1} \times q_{c_2} - q_{c_1} \times q_{b_2}, \quad (13)$$

and

$$q_{a_4} = -q_{a_3}, q_{b_4} = -q_{b_3}, q_{c_4} = -q_{c_3}, q_{d_4} = -q_{d_3}. \quad (14)$$

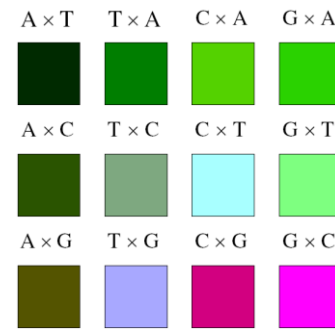


Figure 1. The colors corresponding to the different combination of the multiplied quaternion numbers.

In addition to the real part of the calculated quaternion number, the rest imaginary parts can provide further more mismatching information between the two sequences. A color image in which the R, G, and B components are respectively corresponding to the three imaginary parts $\hat{i}, \hat{j}, \hat{k}$ in the quaternion number can be generated.

First, the possible values of the three imaginary parts are determined and listed in Table I. Second, for each component, the finite values are normalized into the monotone pixel values between 0 and 255.

There are six, five, and three possible values for the \hat{i}, \hat{j} , and \hat{k} imaginary parts, respectively. Each value in each imaginary part can be assigned to as a grayscale value. Therefore, three grayscale images for the R, G, and B components can be generated. By combining the three components, different colors corresponding to the various mismatching conditions between two nucleotides can be obtained. Figure 1 shows the different colors and the corresponding mismatching conditions. According to the color assignments in Fig. 1, a 2-D pattern $f_{Im}(x,y)$ that reflects the imaginary parts of the correlation results can be generated. The image $f_{Im}(x,y)$ is similar to the image $f_{Re}(x,y)$ that reflects

the real parts of the correlation results. However, each pixel represents one of the different mismatching conditions.

D. Complexity Analysis

General performance measurements of an algorithm are the time/computation and space complexity. By definition, the correlation of two sequences $x_1[n]$ and $x_2[n]$ of size N needs N^2 multiplication and $N-1$ addition if the two sequences are of length N . The time complexity of the original correlation is $O(N^2)$. In DSP theories, the discrete Fourier transform (FT) is used to accelerate the speed of correlation (the correlation theorem). That is,

$$r_{1,2}[\xi] = \sum_{l=0}^{n-1} x_1[n]x_2^*[l + \xi] = \text{FT}^{-1}\{X_1^*[\kappa]X_2[\kappa]\}, \quad (15)$$

where $X_1[\kappa] = \text{FT}\{x_1[n]\}$ and $X_2[\kappa] = \text{FT}\{x_2[n]\}$ are the FTs of two sequences $x_1[n]$ and $x_2[n]$, respectively. In Eq. (15), the time complexity depends on the FT. Because the time complexity of the FT is $O(N^2)$ and time complexity of multiplication is $O(N)$, the time complexity of correlation operation is $O(N^2)$. With the fast FT algorithm, the time complexity can be improved to become $O(M\log_2 N)$.

Let FT_Q and FT_Q^{-1} denote the quaternion Fourier transform (QFT) and inverse QFT, respectively. From the Hypercomplex Wiener-Khinchine theorem [23], Eq.(15) becomes:

$$r_{1,2}[\xi] = \text{FT}_Q^{-1}\{X_1[\kappa]X_{2\parallel}^*[\kappa]\} + \text{FT}_Q\{X_1[\kappa]X_{2\perp}^*[\kappa]\}, \quad (16)$$

where $X_1[\kappa] = \text{FT}_Q\{x_1[n]\}$, $X_2[\kappa] = \text{FT}_Q\{x_2[n]\}$, $X_{2\parallel}[\kappa] \parallel \hat{\mu}$, and $X_{2\perp}[\kappa] \perp \hat{\mu}$. Note that $\hat{\mu}$ is the unit pure quaternion and it is referred to as the eigen axis, which represents the direction in the 3-D space of imaginary part of a quaternion. A QFT can be implemented by two ordinary FTs [29]. That is,

$$r_{1,2}[\xi] = \text{FT}^{-1}\{X_{1v}[\kappa]X_{2u}^*[\kappa] - X_{1u}[-\kappa]X_{2v}^*[-\kappa]\}, \quad (17)$$

Where $x_u[n] = x_a[n] + ix_b[n]$, $x_v[n] = x_c[n] - ix_d[n]$. Therefore, the time complexity of quaternion correlation can be significantly reduced to $O(M\log_2 N)$.

Regarding to space complexity, the most memory-consumable is the storage of numerical data for DNA sequences. If the correlation is calculated by using the QFT, it needs additional memory space to store data of frequency domain in the computational process. Therefore, this method trades the memory space for efficiency.

III. EXPERIMENTAL RESULTS

In computer simulation, computer-generated random sequences and real DNA sequences are used to perform the quaternion correlation. Consider a random quaternion sequence $s_{x1}[n]$ and let $s_{x2}[n]$ denote the prefix (first eight quaternion numbers) of the sequence $s_{x1}[n]$. Following the cross-correlation operation, the cross-correlation result is also a quaternion number sequence. Figure 2 shows the real part of the correlation results, which distribute over the eight discrete levels. Actually, there are nine situations in the correlation

results from the matching counts being zero to eight. By observing the coefficients, it shows that each level differs by $4/3$. Therefore, the real parts of cross-correlation coefficients are relative to the matching counts between two sequences. There is a maximum correlation at the position zero because this is the exactly matching position.

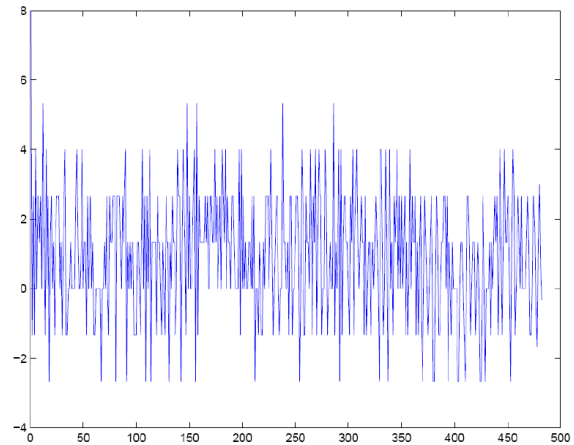


Figure 2. The real-part correlation coefficients of two quaternion sequences $s_{x1}[n]$ and let $s_{x2}[n]$.

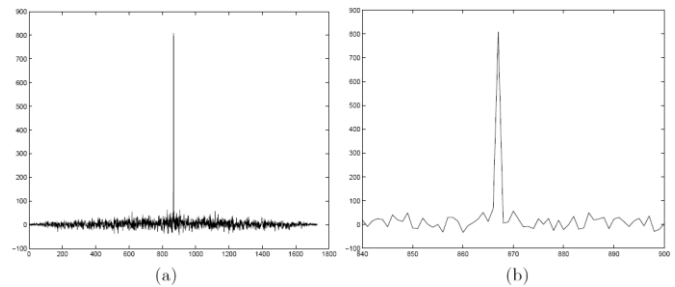


Figure 3. The real-part correlation coefficients of the sequences for the TGFA genes of Human and Mouse. (a) The highest correlation peak appears at the best-matching position; (b) The detailed center region.

Three DNA sequences retrieved from the web-based databases of National Center for Biotechnology Information (NCBI) [30] are then used to test the proposed method. Each sequence has an accession number for identification. First of all, consider two sequences from highly similar genes: the human TGFA sequence (Accession: K03222, 867 bp) and the mouse TGFA sequence (Accession: BC003895, 1024 bp). The DiHydro Folate Reductase (DHFR) gene (Accession: L26316, 1042 bp) from a mouse is also considered. Figure 3(a) shows the quaternion cross-correlation result of the two TGFA genes. A correlation peak with value 807 appearing at the position $\xi = 867$ represents that there exists large similarity (822 identical base pairs) and the best matching position of two sequences can be obtained. Figure 3(b) shows a detailed region from $\xi = 840$ to 900. The correlation values at other positions are much smaller than the peak value. Therefore, it is verified that the proposed method can determine the best global matching position of two sequences. On the other hand, consider the cases of base-pair deletion and insertion, which commonly happens in DNA sequences. To investigate the effects, the sequence of mouse TGFA gene is modified and then used to determine the correlation result. Figure 4(a) and 4(c) show the

correlation results when 1-bp deletion and insertion happen in the center position of the TGFA gene, respectively. As shown in Figs. 4(b) and 4(d) for details, two half values of the original correlation peak value appear at the deletion/insertion positions.

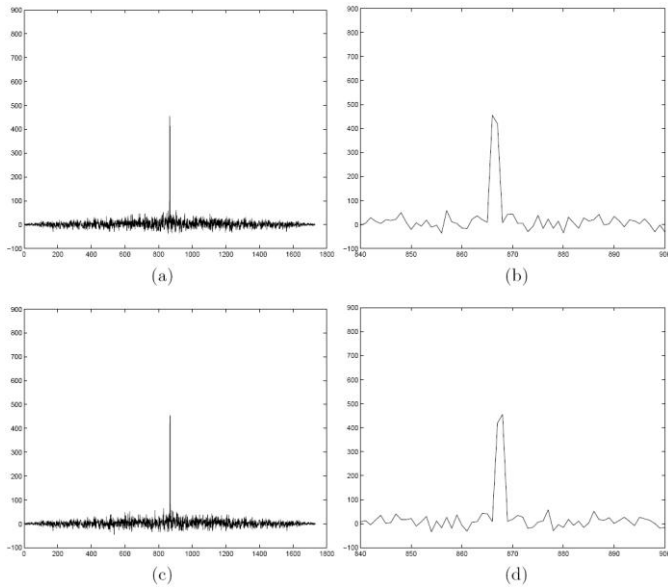


Figure 4. The correlation results for (a) 1-bp deletion in the center position of the Human TGFA sequence; (b) the detailed center region in (a); (c) 1-bp insertion in the center position of the Human TGFA sequence; (d) the detailed center region in (c).

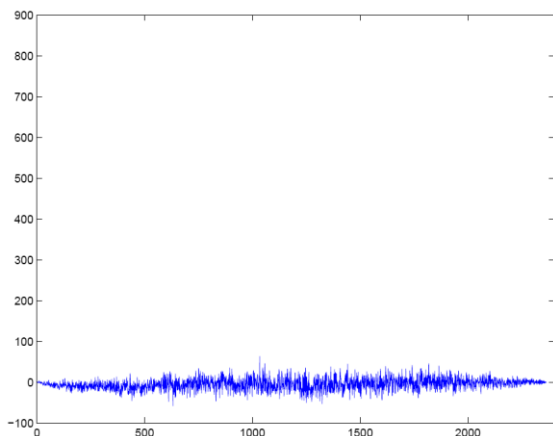


Figure 5. The real-part of cross-correlation coefficients of the sequences for the Human TGFA and Mouse DHFR genes.]

According to the peak value in the correlation result, the number of matched base pairs can be estimated by the use of Eq.(7). If the number of insertion/deletion increases, the corresponding correlation peak values decrease. The partial matching positions can still be detected from the decreased correlation peaks. Therefore, the deletion or insertion in sequences can be estimated from the correlation result. Figure 5 shows the correlation result of two quite different (human TGFA and mouse DHFR) genes. There is no significant peak value among all the real-part coefficients. It verifies that the similarity between these two gene sequences is small.

The sequence matching results obtained from quaternion correlation are compared with the well-known BLAST method [31]. Figure 6 shows the top 34 sequences retrieved by the use of BLAST when using the Human TGFA gene (Accession: K03222, 867 bp) as the query sequence. The sequences for the first 10 high scores (excluding the query sequence itself) are used to perform the quaternion cross-correlation with the query sequence. Each cross-correlation result shows a correlation peak at the best-matching position. Table III summarizes the correlation results and BLAST scores for comparison. The peak values almost follow the trend of BLAST scores. Since BLAST is designed specifically for local alignment of sequences, the small difference between the scores and peak values is reasonable.

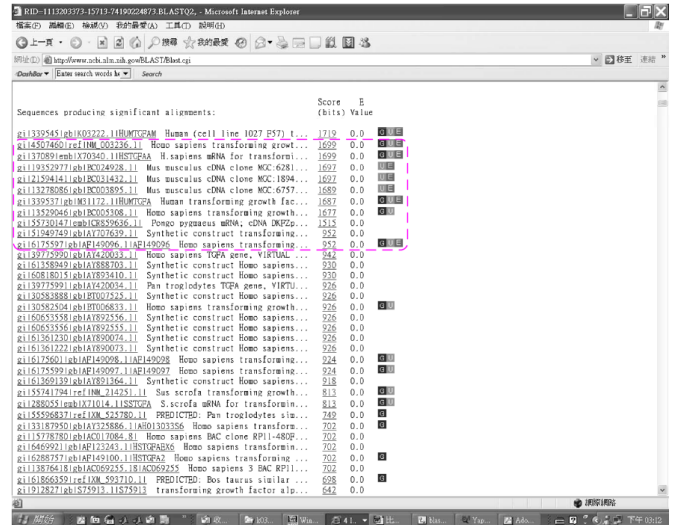


Figure 6. The query result of the Human TGFA gene using the Nucleotide-nucleotide BLAST (blastn) tool.]

TABLE III. THE CORRELATION PEAK VALUES AND POSITIONS AND CORRESPONDING TOP TEN SCORES WHEN USING THE BLASTN TO QUERY THE HUMAN TGFA SEQUENCE IN NCBI DATABASE.

Sequence information	Position	Peak value	Blast score
Homo sapiens TGFA, K03222, 867 bp	867	867	1719
Homo sapiens TGFA, NM.003236, 4119 bp	4122	860	1699
Homo sapiens mRNA for TGFA, X70340, 4119 bp	4122	860	1699
Mus musculus cDNA clone, BC024928, 1097 bp	1027	862	1697
Mus musculus cDNA clone, BC031432, 1109 bp	1032	862	1697
Mus musculus cDNA clone, BC003895, 1042 bp	1035	860	1689
Homo sapiens TGFA, M31172, 867 bp	867	862	1687
Homo sapiens TGFA, BC005308, 1254 bp	889	739	1677
Pongo pygmaeus (orangutan), CR859636, 4135 bp	4019	831	1515
Synthetic construct TGFA, AY707639, 600 bp	634	492	952
Homo sapiens TGFA, AF149096, 782 bp	811	470	952

The local alignment results of two DNA sequences are derived from the real parts of the products during the cross-correlation operation. A 2-D pattern is generated by the use of Eq.(9). Figure 7(a) show the local alignment result for human and mouse TGFA genes. During the cross-correlation operation, the product of two quaternion numbers is not zero only when two sequences overlap. The non-zero products form a parallelogram in the rectangular pattern. A horizontal line appears at the index of correlation $\xi = 867$, which corresponds to the peak correlation result shown in Fig. 7(b). In addition to the cross-correlation values, which represent the global matching information of two sequences,

the local matching information can be observed and measured in the parallelogram. To demonstrate the capability of the proposed quaternion correlation method, Figs. 7(c), 7(e), and 7(g) show the alignment results when 70 bp deletion, insertion, and substitution occur in one of the two sequences. For the cases of deletion and insertion, the horizontal line breaks into two parts, which are shifted horizontally by 70 bp. For the case of substitution, part of the line corresponding to the substituted nucleotides disappears. Figs. 7(d), 7(f), and 7(g) show that the corresponding correlation peaks appear accordingly. In addition to the peak values, the local matching positions can also be directly observed from the 2-D pattern. Compared with the 1-D correlation result, obviously, the 2-D pattern provides more information on the local matching result.

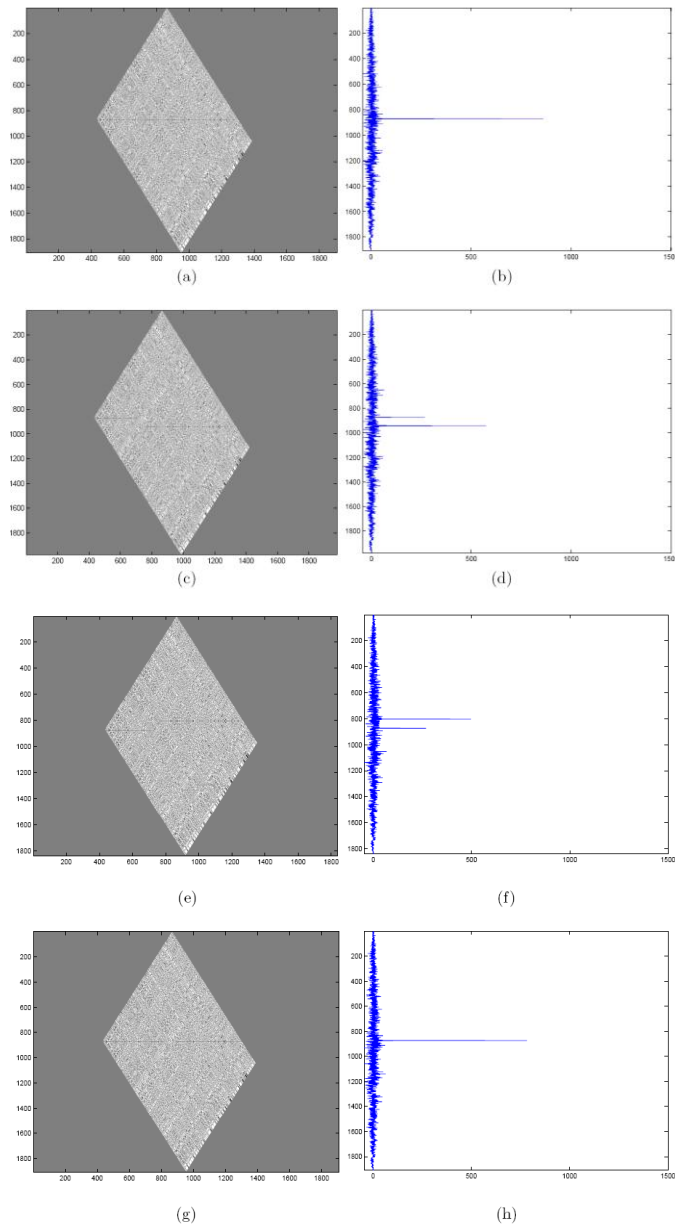


Figure 7. (a) Local matching results obtained from the cross-correlation operation of two quaternion sequences; (b) Corresponding 1-D cross-correlation result of the 2-D pattern shown in (a);(c) 70 bp deletion in one of the sequences;(d) Corresponding 1-D cross-correlation result of the 2-D

pattern shown in (c);(e) 70 bp insertion in one of the sequences;(f) Corresponding 1-D cross-correlation result of the 2-D pattern shown in (e);(g) 70 bp substitution in one of the sequences;(h) Corresponding 1-D cross-correlation result of the 2-D pattern shown in (g).

Finally, Fig. 8 shows the 2-D color pattern in which the mismatching information can be directly observed. According to Fig. 8, four kinds of mismatching (AT, TA, CG, and GC) are the major parts. More information can be observed by examining the detailed parts in this pattern.

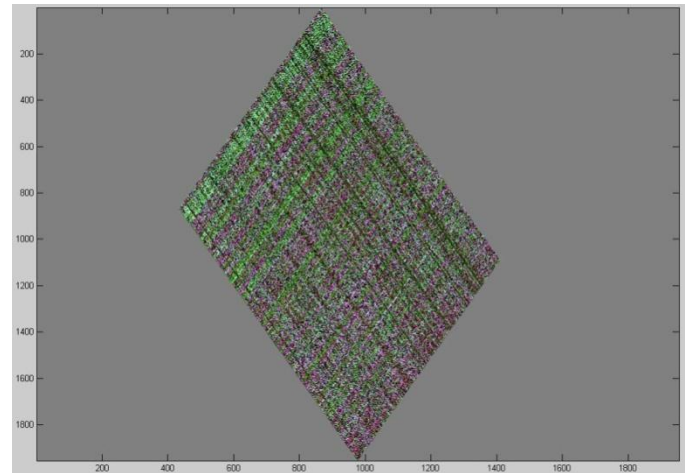


Figure 8. The 2-D pattern corresponding to the imaginary parts of quaternion correlation results of the two sequences for the Human TGFA and Mouse DHFR genes.

IV. CONCLUSION

In this study, the quaternion correlation based on quaternion number systems is proposed for DNA sequence representation and alignment. From the cross-correlation result of quaternion-number sequences, two DNA sequences can be compared in a pair-wise mode. The peak value of real part of the correlation result corresponds to the globally best-matching position of two similar sequences. On the other hand, the 2-D image obtained from the product terms in the cross-correlation operation can provide more information on local alignment of two DNA sequences. For the deletion or insertion happening in the sequences, they can be discriminated by analyzing the correlation results. Moreover, a color 2-D image can also be generated to visualize the mismatching conditions of two DNA sequences. The simulation results show that the proposed method is of promising potential in bioinformatics. Future work will focus on extracting more information and relationships between two sequences from the generated 2-D pattern.

ACKNOWLEDGMENT

This work is partly supported by National Science Council, Taiwan, under the contract number NSC 100-2628-E-224-002-MY2.

REFERENCES

- [1] C. Zhang and A.K. Wong, "A genetic algorithm for multiple molecular sequence alignment," *Comput. Appl. Biosci.*, vol. 13, pp. 565–581, 1997.
- [2] W. Choe, O.K. Ersoy, and M. Bina, "Neural network schemes for detecting rare events in human genomic DNA," *Bioinformatics*, vol. 16, pp. 1062–1072, 2000.

- [3] L. Hirschman, J.C. Park, J. Tsujii, L. Wong, and C.H. Wu, "Accomplishments and challenges in literature data mining for biology," *Bioinformatics*, vol. 18, pp. 1553–1561, 2002.
- [4] P.D. Cristea, "Large scale features in DNA genomic signals," *Signal Processing*, vol. 83, no. 4, pp. 871–888, 2003.
- [5] <http://www.expasy.org/links.html>
- [6] <http://www.ensembl.org/index.html>
- [7] <http://www.ncbi.nlm.nih.gov/Entrez/>
- [8] D. Anastassiou, "Genomic signal processing," *Signal Processing Magazine*, vol. 18, pp. 8–20, 2001.
- [9] J. Astola, E. Dougherty, I. Shmulevich, and I. Tabus, "Genomic signal processing," *Signal Processing*, vol. 83, no. 4, pp. 691–694, 2003.
- [10] A.V. Oppenheim, R.W. Schaffer, and J.R. Buck, *Discrete-Time Signal Processing*, 2nd Edition, Chapter 1, Prentice Hall International, Inc., 1999.
- [11] S.-Y. Hsieh, C.-W. Huang, and H.-H. Chou, "A DNA-based graph encoding scheme with its applications to graph isomorphism problems," *Applied Mathematics and Computation*, vol. 203, no. 2, pp. 502–512, September 2008.
- [12] S.-Y. Hsieh and M.-Y. Chen, "A DNA-based solution to the graph isomorphism problem using Adleman-Lipton model with stickers," *Applied Mathematics and Computation*, vol. 197, no. 2, pp. 672–686, April 2008.
- [13] W. Wang and D.H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Transactions on Signal Processing*, vol. 50, pp. 628–634, 2002.
- [14] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *Computer Applications in the Biosciences*, vol. 13, pp. 263–270, 1997.
- [15] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," *Bioinformatics*, vol. 16, no. 12, pp. 1073–1081, 2000.
- [16] Y. Magarshak, "Quaternion representation of RNA sequences and tertiary structures," *Biosystems*, vol. 30, no. 1–3, pp. 21–29, 1993.
- [17] T. Bulow and G. Sommer, "Hypercomplex signals— A novel extension of the analytic signal to the multidimensional case," *IEEE Transactions on Signal Processing*, vol. 49, no. 11, pp. 2844–2852, 2001.
- [18] J.-J. Shu and L.S. Ow, "Pairwise alignment of the DNA sequence using hypercomplex number representation," *Bulletin of Mathematical Biology*, vol. 66, no. 5, pp. 1423–1438, 2004.
- [19] Y. Li and J.-J. Shu, "Cross-correlation of DNA sequences using hypercomplex number encoding," 2006 International Conference on Biomedical and Pharmaceutical Engineering (ICBPE 2006), pp. 453–458, 11–14 Dec. 2006.
- [20] J.-J. Shu and Y. Li, "Hypercomplex cross-correlation of DNA sequences," *Journal of Biological Systems*, vol. 18, no. 4, pp. 711–725, 2010.
- [21] I.L. Kantor and A.S. Solodovnikov, *Hypercomplex Number: An Elementary Introduction to Algebras*, New York: Springer-Verlag, 1989.
- [22] S.J. Sangwine and T.A. Ell, "Color image filter based on hypercomplex convolution," *IEE Proceedings of Vision, Image Signal Processing*, vol. 147, no. 2, pp. 89–93, 2000.
- [23] T.A. Ell and S.J. Sangwine, "Hypercomplex Wiener-Khinchin theorem with application to color image correlation," *IEEE International Conference on Image Processing*, vol. 2, pp. 792–795, 2000.
- [24] K. Ueda, S.-I. Takahashi, "Digital filters with hypercomplex coefficients," 1993 IEEE International Symposium on Circuits and Systems, vol. 1, pp. 479–482, 1993.
- [25] <http://www.wikipedia.org/>
- [26] H.T. Chang, N.W. Lo, W.C. Lu, and C.J. Kuo, "Visualization of DNA sequences by use of three-dimensional trajectory," *The First Asia-Pacific Bioinformatics Conference*, vol. 19, pp. 81–85, Australia, Feb. 2003.
- [27] H.T. Chang, "DNA sequence visualization," Chapter 4 in *Advanced Data Mining Technologies in Bioinformatics*, pp. 63–84, Edited by Dr. H.-H. Hsu, Idea Group, Inc., 2006.
- [28] N.-W. Lo, H.T. Chang, S.W. Xiao, and C.J. Kuo, "Global visualization of DNA sequences by use of three-dimensional trajectories," *Journal of Information Science and Engineering*, vol. 23, no. 6, pp. 1723–1736, Nov. 2007.
- [29] S.-C. Pei, J.J. Ding, and J.H. Chang, "Efficient implementation of quaternion Fourier transform, convolution, and correlation by 2-D Complex FFT," *IEEE Transactions on Signal Processing*, vol. 49, no. 11, pp. 2783–2797, Nov. 2001.
- [30] <http://www.ncbi.nlm.nih.gov/>
- [31] <http://www.ncbi.nlm.nih.gov/BLAST/>

AUTHORS PROFILE

Hsuan T. Chang received his B.S. degree in Electronic Engineering from National Taiwan University of Science and Technology, Taiwan, in 1991, and M.S. and Ph.D. degree in Electrical Engineering (EE) from National Chung Cheng University (NCCU), Taiwan, in 1993 and 1997, respectively. He was a visiting researcher at Laboratory for Excellence in Optical Data Processing, Department of Electrical and Computer Engineering, Carnegie Mellon University, from 1995 to 1996. Dr. Chang was an assistant professor in Department of Electronic Engineering in Chien-Kuo University of Technology, Changhua, Taiwan, from 1997 to 1999, an assistant professor in Department of Information Management, Chaoyang University of Technology, Wufeng, Taiwan, from 1999 to 2001, and an assistant professor and associate professor in EE Department of National Yunlin University of Science and Technology (YunTech), Douliu, Taiwan, from 2001 to 2002 and 2003 to 2006, respectively. He served as Chairman of Graduate Institute of Communications Engineering of Yuntech from 2008–2011. He currently is a full professor in EE Department of YunTech, Chairman of Graduate School of Engineering Technology and Science, and Deputy Dean of College of Engineering, YunTech. He was also an adjunct assistant professor in Graduate Institute of Communications Engineering of NCCU from 2000 to 2003. Dr. Chang was a visiting scholar in Institute of Information Science, Academia Sinica, Taiwan and in EE department, University of Washington, Seattle USA from 2003/7 to 2003/9 and 2007/8 to 2008/3, respectively. Dr. Chang's interests include image/video analysis, optical information processing/computing, medical image processing, and human computer interface. He has published more than 180 journal and conference papers in the above research areas. He was the recipient of the visiting research fellowship from Academia Sinica, Taiwan in 2003, and the excellent research award for young faculty in NYUST in 2005. He also received Outstanding Paper Award from 2009 MATLAB/Simulink Tech Forum Call for Papers in 2009, Taiwan. He served as the reviewer of several international journals. He served as the conference chair of 2005 Workshop on Consumer Electronics and Signal Processing held in Taiwan and was an invited speaker, session chair, and program committee in various domestic and international conferences. Dr. Chang is Senior Member of Institute of Electrical and Electronics Engineers (IEEE), Senior Member of Optical Society of America (OSA), International Society for Optical Engineering (SPIE), a member of International Who's Who (IWW), Taiwanese Association of Consumer Electronics (TACE), Asia-Pacific Signal and Information Processing Association (APSIPA), and The Chinese Image Processing and Pattern Recognition (IPPR) Society.

Chung J. Kuo received BS and MS degree in Power Mechanical Engineering from National Tsinghua University, Taiwan, in 1982 and 1984, respectively, and PhD degree in Electrical Engineering (EE) from Michigan State University (MSU) in 1990. He joined EE Department of National Chung Cheng University (NCCU) in 1990 as an associate professor and then became a full professor in 1996. He was the chairman of Graduate Institute of Communications Engineering of NCCU between 1999 and 2002. Dr. Kuo was a visiting scientist at Opto-Electronics & System Lab, Industrial Technology Research Institute in 1991 and IBM T.J. Watson Research Center from 1997 to 1998 and a consultant to several international/local companies. He was the Director of RD Center of Components Business Group (CPBG), Delta Electronics, Inc. from 2003 to 2004. In 2004, Dr. Kuo became the Senior Director of Magnetism and Microwave Business Unit of CPBG, Delta Electronics, Inc. Dr. Kuo currently is consultants of two private companies. Dr. Kuo interests in image/video signal processing, VLSI signal processing, and photonics. He is the co-director of the Signal and Media (SAM) Labs., NCCU. Dr. Kuo received the Distinguished Research Award from National Chung Cheng University in 1998, Overseas Research Fellowship from National Science Council (NSC) in 1997, Outstanding Research Award from College of Engineering, NCCU in 1997, Medal of Honor from NCCU in 1995, Research Award from NSC for consecutive 11 times, EE Fellowship from MSU in 1989, and Outstanding Academic Achievement Award from MSU in

1987. He was a guest editor for three special sections of Optical Engineering and 3D Holographic Imaging (to be published by John, Wiley and Sons) and an invited speaker and program committee chairman/member for several international/local conferences. He also serves as an Associate Editor of IEEE Signal Processing Magazine and President of SPIE Taiwan Chapter (1998-2000). Dr. Kuo is a senior member of IEEE and a member of Phi Kappa Phi, Phi Beta Delta, OSA, and SPIE and listed in Who's Who in the World.

Neng-Wen Lo received his bachelor's degree in Physics from the Tunghai University in 1986 and Ph.D. degree in Biochemistry from the State University of New York at Buffalo, USA, in 1997. He was a research fellow at the Oncology Center of Johns Hopkins School of Medicine in Maryland,

USA, from 1997 to 1999. In 1999, he returned to Taiwan and joined the Department of Animal Science and Biotechnology at Tunghai University. He is currently an Associate Professor and Head of the Agriculture Extension Center, College of Agriculture. Dr. Lo's research interest is in Computational Biology and in Reproduction Biology. He has published more than 40 journal and conference papers. He was the chief editor of Tunghai Journal in 2003 and ever served as referees for several journals. Dr. Lo is a founding member of the Bioinformatics Society in Taiwan. He is also a member of the Society for the Study of Reproduction and a permanent member of Chinese Society of Animal Science.

Wei-Z.LvBiography is not available.

Evaluation of Regressive Analysis Based Sea Surface Temperature Estimation Accuracy with NCEP/GDAS Data

Kohei Arai¹

Graduate School of Science and Engineering
Saga University
Saga City, Japan

Abstract—In order to evaluate the skin surface temperature (SSST) estimation accuracy with MODIS data, 84 of MODIS scenes together with the match-up data of NCEP/GDAS are used. Through regressive analysis, it is found that 0.305 to 0.417 K of RMSE can be achieved. Furthermore, it also is found that band 29 is effective for atmospheric correction (30.6 to 38.8% of estimation accuracy improvement). If single coefficient set for the regressive equation is used for all the cases, SSST estimation accuracy is around 1.969 K so that the specific coefficient set for the five different cases have to be set.

Keywords- Thermal infrared radiometer; Skin sea surface temperature; Regressive analysis; Split window method; NCEP/GDAS; MODTRAN; Terra and AQUA/MODIS.

I. INTRODUCTION

The required skin sea surface temperature estimation accuracy is better than 0.25K for radiation energy budget analysis, global warming study and so on [1]. Skin sea surface temperature (SSST) is defined as the temperature radiation from the sea surface (approximately less than 20 μ m in depth from the surface) and is distinct with the Mixed layer sea surface temperature (MSST) based on the temperature radiation from the sea surface (about 10 m in depth from the surface) and also is different from the Bulk sea surface temperature (BSST) which is based on the temperature radiation from just below the skin [2]. In order to estimate SSST, (a) atmospheric influences which are mainly due to water vapor followed by aerosols for the atmospheric window channels [3], (b) cloud contamination, (c) emissivity changes mainly due to white caps, or forms followed by limb darkening due to changes of path length in accordance with scanning angle changes should be corrected [4]-[8].

One of the atmospheric corrections is split window method which is represented by Multi Channel Sea Surface Temperature (MCSST) estimation method [9]. In the MCSST method, 10 μ m of atmospheric window is split into more than two channels. The atmospheric influences are different among the split channels so that it is capable to estimate atmospheric influence by using the difference. Through a regressive analysis between the estimated SSST with acquired channels of satellite onboard thermal infrared radiometer (TIR) data and the corresponding match-up truth data such as buoy or shipment data (Bulk temperature) with some errors, all the

required coefficients for the regressive equation are determined. Thus SSST is estimated with the regressive equation if the newly acquired TIR data is put into the regressive equation. On the other hand, the method for improvement of SSST estimation accuracy by means of linearized inversion of radiative transfer equation (RTE) is proposed [10] and also the method for solving RTE more accurately based on iterative method is proposed [11],[12]. RTE is expressed with Fred-Holm type of integral equation with a variety of parameters. Such RTE can be solved with linear and/or Non-Linear inversion methods. Integral equation can be linearized then RTE can be solved by using linear inversion method and also can be solved iteratively. The former and the later are called Linearized inversion and Non-Linear iteration, respectively.

In accordance with MODTRAN [13], the following 5 cases of the ocean areas (Latitude) and the seasons are selected, (a)Sub-Arctic Summer, (b)Sub-Arctic Winter, (c)Mid-Latitude Summer, (d)Mid-Latitude Winter, (e)Tropic. The regressive analysis is made by using a match-up data set of MODIS data and NCEP/GDAS (Global Data Assimilation Model [14], 1 degree mesh data of air temperature, relative humidity, ozone, cloud, precipitable water for the sphere with the altitude ranges from 1000 to 100 hPa) data. The regressive error in terms of Root Mean Square Error (RMSE) and regressive coefficients are calculated.

Firstly, major specification of MODIS is introduced together with atmospheric characteristics such as transparency, water vapor profile and aerosol profile derived from MODTRAN. Secondly, the method for regressive analysis is described followed by the procedure of the preparation of match-up data derived from NCEP/GDAS and MODIS data together with cloud masking. Thirdly, the results from the regressive analysis are shown followed by the results from the comparative study on regressive equations. Finally, major conclusions are discussed.

II. PROPOSED METHOD

A. MODIS onboard Terra and Aqua satellites

MODIS/TIR onboard Terra and AQUA satellites is moderate spatial resolution (IFOV=1k m) of thermal infrared radiometer with the swath width of 2400 km and consists of 3

channels of radiometer which covers the wavelength region shown in Table 1.

B. Atmospheric Model Used

The atmospheric transmittance for the wavelength region for the typical atmospheric models in accordance with MODTRAN 4.0, Tropic, Mid-Latitude Summer and winter, Sub-Arctic Summer and winter are shown in Fig.1.

TABLE I. WAVELENGTH COVERAGE OF MODIS

Cloud Properties	29	8.400 - 8.700	9.58(300K)	0.05
Surface/Cloud Temperature	31	10.780 - 11.280	9.55(300K)	0.05
	32	11.770 - 12.270	8.94(300K)	0.05

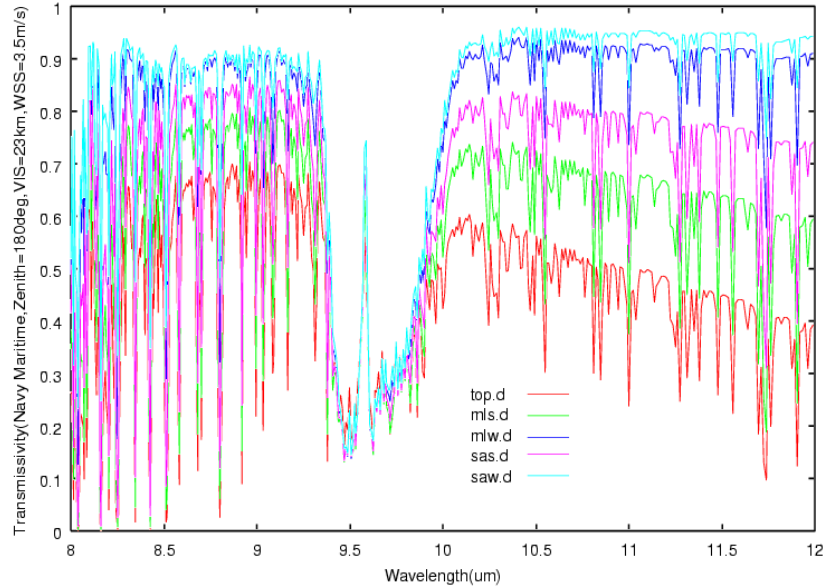


Figure 1 Transmittance for the typical atmospheric models

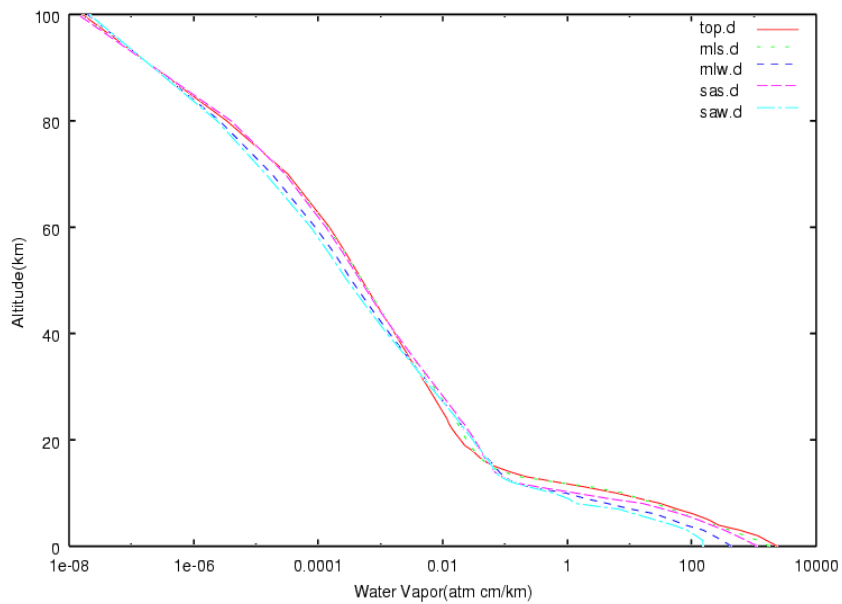


Fig2 The vertical profile of the water vapor for the typical atmospheric models.

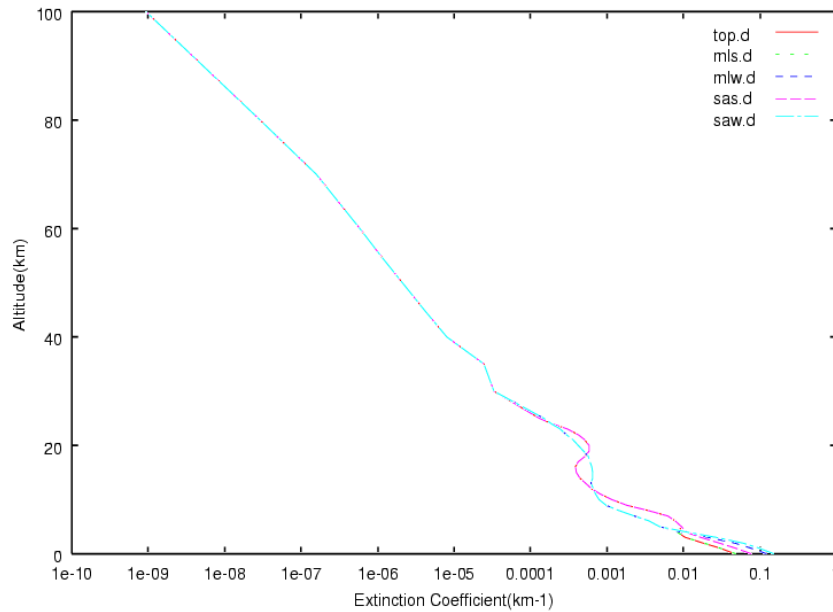


Fig.3 The vertical profile of the aerosol extinction coefficient for the typical atmospheric models

In this calculation, aerosol type and meteorological range are set to rural and 23 km, respectively while the observation angle is set at 0 degree of zenith angle, namely, nadir looking situation. The vertical profiles of water vapor content in the atmosphere (in unit of atm cm/km) and aerosol extinction coefficient (in unit of km⁻¹) are shown in Fig.2 and 3, respectively.

The Tropic shows lowest transmittance followed by Mid-Latitude Summer, Sub-Arctic Summer, Mid-Latitude Winter and Sub-Arctic Winter while the Tropic shows largest water vapor content in the atmosphere followed by Mid-Latitude Summer, Sub-Arctic Summer, Mid-Latitude Winter and Sub-Arctic Winter.

The most dominant factor for atmospheric influence in this wavelength region is water vapor followed by aerosol so that vertical profiles have to be clarified. Relatively good MODIS data for the Sub-Arctic Winter could not be found so that (1) Tropic, (2) Mid-Latitude Summer, (3) Mid-Latitude Winter and (4) Sub-Arctic Summer were selected for the analysis.

C. Regressive Analysis

The following simple linear regressive equation between NCEP/GDAS derived SSST and the physical temperature estimated with MODIS Level 1B product of data is assumed,

$$SSST=C_1+C_2*T_{30}+C_3*(T_{30}-T_{31})+C_4*(\sec(\theta)-1)*(T_{30}-T_{31})+C_5*(T_{30}-T_{29})*(\sec(\theta)-1)*(T_{30}-T_{29}) \quad (1)$$

where T_i is the brightness temperature of the channel i , θ is the solar zenith angle. Through a regressive analysis, C_i for each case is estimated as well as regression error, Root Mean Square Error (RMSE) which is corresponding to the SSST estimation error. In this connection, the swath width of MODIS is 60 km while the grid size of NCEP/GDAS is 1 degree so that the averaged T_i over full scene of MODIS (2400 km) is calculated and put into the equation (1) together with the linearly interpolated NCEP/GDAS data at the scene center of the MODIS data which become SSST in the equation (1). In accordance with ATBD 25 (SST algorithm [15]), SST is estimated with the following equation,

$$SST = a_0+a_1T_1+a_2(T_1-T_2)T_b+a_3(\sec(\theta) -1) \quad (2)$$

where T_b is MCSST [9]. The proposed regressive equation uses band 29 of brightness temperature influenced by water vapor. It is very effective for atmospheric correction.

III. EXPERIMENTS

The relations between MODIS derived SSST and NCEP/GDAS SSST for the typical atmospheric models of Tropic, Mid-Latitude Summer and winter and Sub-Arctic Summer are shown in Fig.4 to 7. In the experiments, 37, 18, 15 and 14 MODIS data which were acquired from 2000 to 2001 were used together with the match-up NCEP/GDAS data.

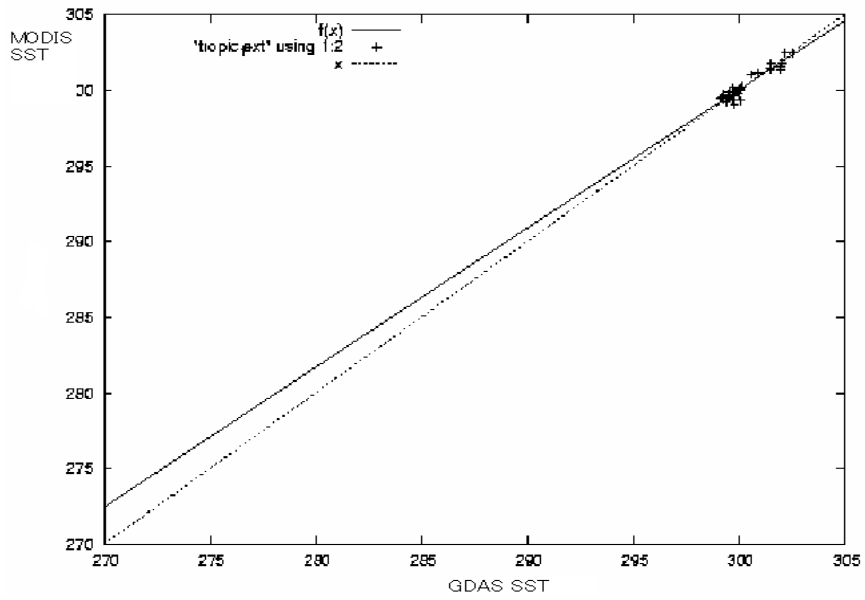


Figure 4 The relation between MODIS derived SSST and NCEP/GDAS SSST for Tropic ocean area.

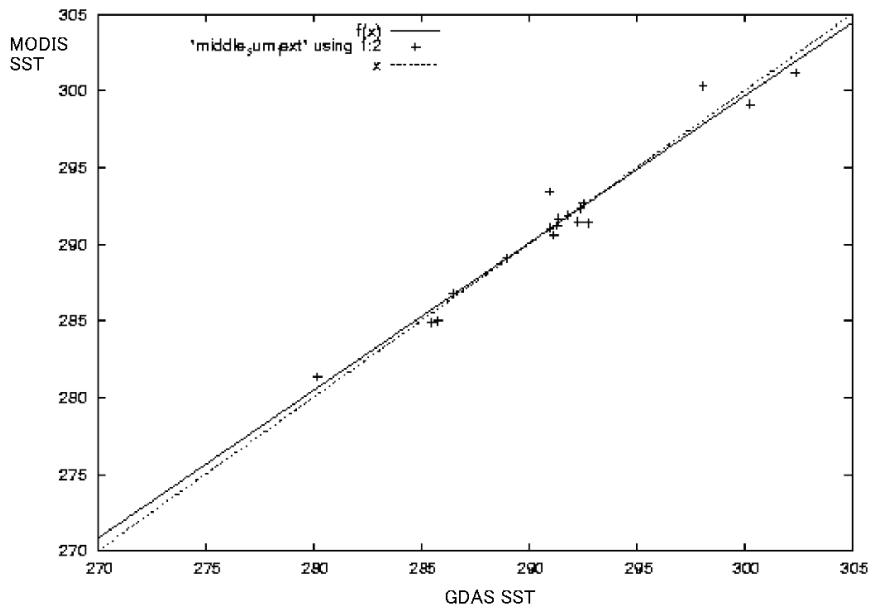


Figure 5 The relation between MODIS derived SSST and NCEP/GDAS SSST for Mid-Latitude Summer of ocean area and season

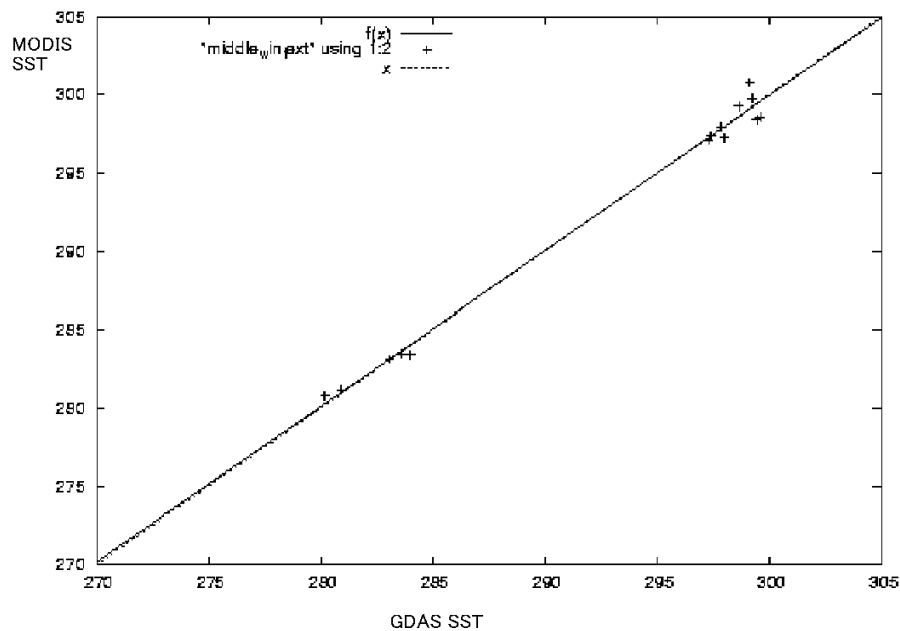


Figure 6 The relation between MODIS derived SSST and NCEP/GDAS SSST for Mid-Latitude Winter of ocean area and season

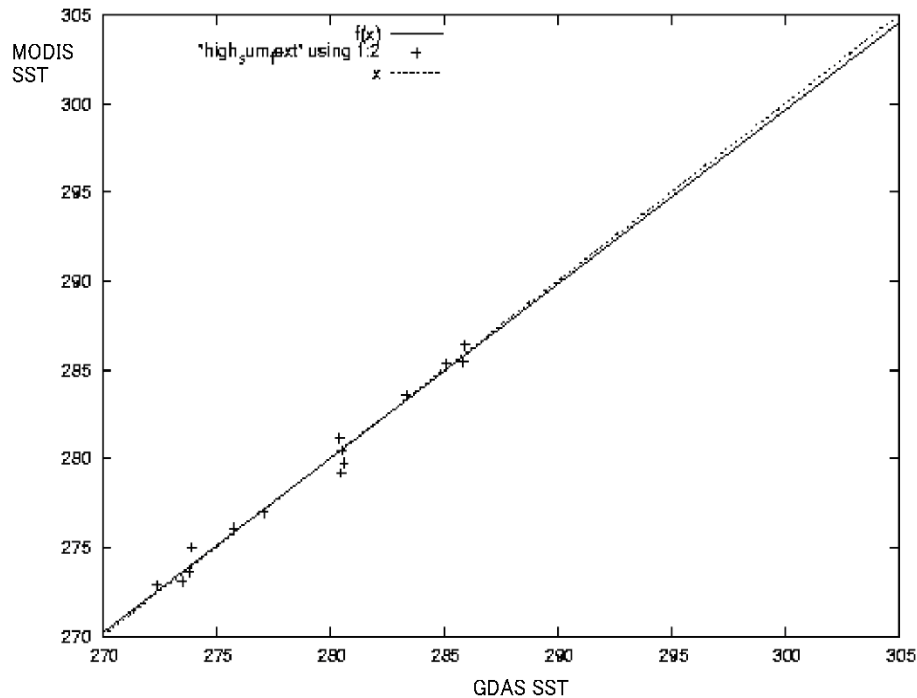


Figure 7 The relation between MODIS derived SSST and NCEP/GDAS SSST for Sub-Arctic Summer of ocean area and season

For all the ocean areas and seasons, in particular, for Tropic ocean area, it may be concluded that MODIS derived SSST is lower than NCEP/GDAS derived SSST for the relatively high SSST portion while MODIS derived SSST is higher than that from NCEP/GDAS for the relatively low SSST portion. For the tropic ocean area, SSST ranges just from 298 K to 303 K while the ranges for the Mid-Latitude Summer and winter as well as Sub-Arctic Summer are 281-302 K, 280-299 K and 272-287 K, respectively.

It is also found that the effectiveness of band 29 on RMSE improvement for Tropic model is greatest followed Mid-

Latitude Summer, Sub-Arctic Summer and Mid-Latitude Winter. This order is the same order of water vapor and aerosol content in the atmosphere as is shown in Fig.3 and 4.

IV. CONCLUSION

It is found that SSST estimation accuracy with MODIS data is better than 0.417 K for all the cases defined above. A comparison of SSST estimation accuracy among the cases is attempted. Through regressive analysis, it is found that 0.305 to 0.417 K of RMSE can be achieved.

Furthermore, it is also found that the effectiveness of exclusion of the band 29 of on RMSE degradation for Tropic model is greatest followed Mid-Latitude Summer, Sub-Arctic Summer and Mid-Latitude Winter. This order is the same order of water vapor and aerosol content in the atmosphere. RMSE of the regressive equation without band 29 ranges from 0.498 to 6.01 K . Thus it is concluded that band 29 is effective to atmospheric correction (water vapor absorption). The effect corresponds to 30.6 to 38.8% of SSST estimation accuracy improvement.

If single set of regressive coefficients for all the cases is used, then SSST estimation accuracy is around 1.969 K so that the regressive equations with the specific five different cases, Tropic, Mid-Lat. Summer and winter and Sub-Arctic Summer and winter have to be used.

REFERENCES

- [1] NASA, Science and mission requirements Working Report, EOS Science Steering Committee Report, Vol.I, 1990.
- [2] Barton,I.J., Satellite-derived sea surface temperatures: Current status, Journal of Geophysical Research, Vol.100, pp.8777-8790, 1995. and Personal correspondence at the 37th COSPAR Congress, Warsaw, Poland, July 2000.
- [3] Scott, N.A. and A.Chedin, A fast line-by-line method for atmospheric absorption computations: the automatized atmospheric absorption atlas, Journal of Applied Meteorology, Vol.20, pp.802-812, 1981.
- [4] Hollinger, J.P, Passive microwave measurements of sea surface roughness, IEEE Trans. on Geoscience and Remote Sensing, Vol.GE-9, No.3, 1971.
- [5] Cox, C. and W.H.Munk, Some problems in optical oceanography, Journal on Marine Research, Vol.14, pp.68-78, 1955.
- [6] Harris, A.R., O. Brown and M.Mason, The effect of wind speed on sea surface temperature retrieval from space, Geophysical Research Letters, Vol.21, No.16, pp.1715-1718, 1994.
- [7] Masuda K., T.Takashima and T.Takayama, Emissivity of pure and sea waters for the model sea surface in the infrared window regions, Remote Sensing Environment, Vol.24, pp.313-329, 1988.
- [8] Watte, P.D., M.R.Allen and T.J.Nightingale, Wind speed effect on sea surface emission and reflection for along track scanning radiometer, Journal of Atmospheric and Oceanic Technology, Vol.13, pp.126-141, 1996.
- [9] McClain, E.P., W.G.Pichel and C.C.Walton, Comparative performance of AVHRR based multi-channel sea surface temperatures, Journal of Geophysical Research, Vol.90, No.C6, pp.11587-11601, 1985.
- [10] Arai K.,K.Kobayashi and M.Moriyama, Method to improve sst estimation accuracy by means of linearized inversion of the radiative transfer equation, Journal of Remote Sensing Society of Japan, Vol.18, No.3, pp.272-279, 1998.
- [11] Arai, K., Y.Terayama, M.Miyamoto and M.Moriyama, SST estimation using thermal infrared radiometer data by means of iterative method, Journal of Japan Society of Photogrammetry and Remote Sensing, Vol.39, No.1, pp.15-20, 2000.
- [12] Arai, K., Influence due to aerosols on SST estimation accuracy for the non-linear iterative method, Journal of Remote Sensing Society of Japan, Vol.21, No.4, pp.358-365, 2001.
- [13] US Air Force Geophysical Laboratory, MODTRAN 3 User Instructions, GL-TR-89-0122, 1996.
- [14] NOAA/NGDC, <ftp://ftp.ngdc.noaa.gov/pub/NCEP/GDAS/>
- [15] NASA/GSFC, <http://picasso.oce.orst.edu/ORSOO/MODIS/code/ATBDsum.html#SST>

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission "A" of ICSU/COSPAR since 2008. He wrote 30 books and published 322 journal papers.

The Modelling Process of a Paper Folding Problem in GeoGebra 3D¹

Muharrem Aktumen
Department of Mathematics
Education
AhiEvrان University
Kirsehir, Turkey

Bekir Kursat Doruk
Department of Mathematics
Education
AhiEvrان University
Kirsehir, Turkey

Tolga Kabaca
Department of Mathematics
Education
Pamukkale University
Denizli, Turkey

Abstract—In this research; a problem situation, which requires the ability of thinking in three dimensions, was developed by the researchers. As the purpose of this paper is producing a modeling task suggestion, the problem was visualized and analyzed in GeoGebra3D environment. Then visual solution was also been supported by algebraic approach. So, the capability of creating the relationship between geometric and algebraic representations in GeoGebra was also presented in 3D sense.

Keywords-component; Modelling; GeoGebra 3D; Paper Folding.

I. INTRODUCTION

There are several studies on modeling the real life situations in GeoGebra [2, 3, 4, 5]. In this research, a problem situation has been suggested and modeling process in GeoGebra has been explained. Zbiek and Conner pointed out, modeling contributes to understand the fore known mathematical concepts thoroughly by demonstrating the applicability of mathematical thoughts to real life, to learn new mathematical concepts, to establish inter disciplinary relations and to both conceptual and operational development of the students studying in modeling processes [6]. Furthermore, the algebraic and geometric representations are needed to be connected in two ways [1,7]. That is, the modeling should present the advantage of understanding how algebraic facts effect the observed situations.

The problem can be described as follows according to the figure 1 and figure 2. Let's call the intersection point of the segment $[KL]$ and the line passing through the point D and parallel to segment $[OM]$ as B and the intersection point of the segment $[DK]$ and the line passing through the point C and parallel to the segment $[KL]$ as A . So, the rectangle $DABC$ and the triangles $\triangle DOA$, $\triangle AKB$, $\triangle BLC$ and $\triangle CMD$ can be obtained, such that the points D and A can be moved dynamically on the segments on which they are located.¹

¹A short summary of this article has been submitted to The International GeoGebra Institute Conference 2012 on 21-23 September 2012 and supported by Ahi Evran University.

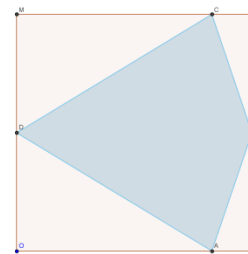


Figure 1: Statement of pre-folding

The problem was stated as “for which locations of the dynamic points A and B , the segments $[OL_t]$ and $[K_tM_t]$ intersect?” in the research (Figure 2). The mathematical concepts related to the solution can be summarized as algebraic approach, line equations, slope, point and vector in three dimensional space and scalar triple product. As a short result, it can be stated that at least one of the points A and B must be in the midpoint of the segments on which they are located after visualization in the environment of GeoGebra 5.0. It is expected that this kind of real situated activities are attractive for the students.

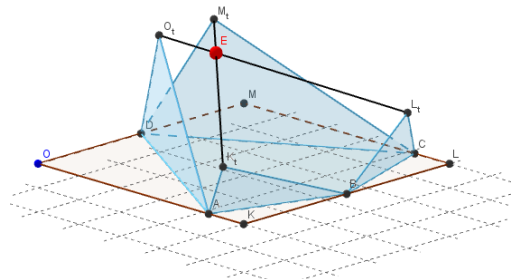


Figure 2: Statement of post-folding

II. UNDERSTANDING THE MODEL VISUALLY

The basic structure of the problem can be constructed as follows where the values of a , b , x_0 and y_0 defined as a slider tool. The points $O(0,0)$, $K(a,0)$, $L(a,b)$, $M(0,b)$ are the corners of the rectangle. The points $A(x_0,0)$, $C(x_0,b)$, $D(0,y_0)$ and $B(a,y_0)$, are dynamic points which are located on the sides of the rectangle and controlled by the sliders dynamically.

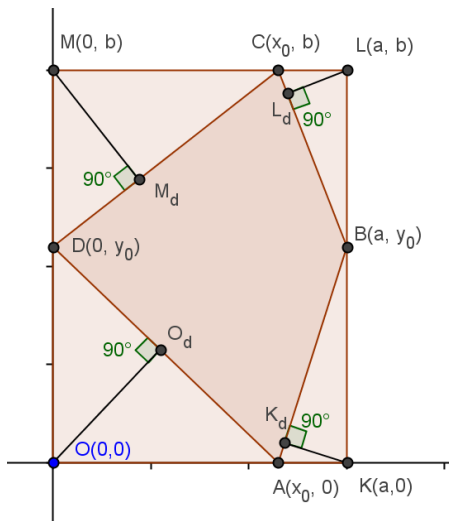


Figure 3: Problem statement

III. ANALYZING THE MODEL ALGEBRAICALLY

After the triangles DOA , AKB , BLC and CMD are folded on the sides DA , AB , BC and DC respectively, the coordinates of the points O_d , K_d , L_d and M_d must be determined.

Finding the coordinate of the point O_d :

The intersection point of the line d_{DA} and the line d_{OO_d} will provide us to determine the coordinates of the point O_d .

Since $m_{d_{DA}} = -\frac{y_0}{x_0}$, by using the fact of the product of the tangents of two perpendicular lines, it can be obtained that $m_{d_{OO_d}} = \frac{x_0}{y_0}$.

Both of the lines' equations can be constructed by using a point which belongs to the line and its tangent.

The equation of the line d_{DA} can be calculated as follows;

$d_{DA} : y = -\frac{y_0}{x_0}(x - x_0) \Rightarrow yx_0 = -y_0x + x_0y_0$. After the revision of equation, following form can be obtained;

$$yx_0 + y_0x = x_0y_0 \quad (1)$$

The equation of the line d_{OO_d} can be calculated as follows;

$d_{OO_d} : y = \frac{x_0}{y_0}x \Rightarrow yy_0 = x_0x$. After the revision of equation, following form can be obtained;

$$yy_0 - x_0x = 0 \quad (2)$$

By finding the common solution of the equation (1) and (2) the coordinates of O_d can be obtained as follows;

$$O_d \left(\frac{x_0 y_0^2}{x_0^2 + y_0^2}, \frac{x_0^2 y_0}{x_0^2 + y_0^2} \right) \quad (3)$$

Finding the coordinates of the point K_d :

The intersection point of the line d_{AB} and the line d_{KK_d} will provide us to determine the coordinates of the point K_d .

Since $m_{d_{AB}} = \frac{y_0}{a - x_0}$, by using the fact of the product of the tangents of two perpendicular lines, it can be obtained that $m_{d_{KK_d}} = -\frac{a - x_0}{y_0}$.

Both of the lines' equations can be constructed by using a point which belongs to the line and its tangent.

The equation of the line d_{AB} can be calculated as follows;

$$d_{AB} : y = \frac{y_0}{a - x_0}(x - x_0) \Rightarrow y(a - x_0) = y_0(x - x_0) \quad \text{After}$$

$$\Rightarrow ya - yx_0 = y_0x - x_0y_0$$

the revision of equation, following form can be obtained;

$$y(a - x_0) - y_0x = -x_0y_0 \quad (4)$$

The equation of the line d_{KK_d} can be calculated as follows;

$$d_{KK_d} : y = -\frac{a - x_0}{y_0}(x - a) \Rightarrow yy_0 = -(a - x_0)(x - a) \quad \text{After}$$

the revision of equation, following form can be obtained;

$$yy_0 + (a - x_0)x = a(a - x_0) \quad (5)$$

By finding the common solution of the equation (4) and (5) the coordinates of K_d can be obtained as follows;

$$K_d \left(\frac{y_0^2 x_0 + a(a - x_0)^2}{y_0^2 + (a - x_0)^2}, \frac{y_0(a - x_0)^2}{y_0^2 + (a - x_0)^2} \right) \quad (6)$$

Finding the coordinates of the point L_d :

The intersection point of the line d_{BC} and the line d_{LL_d} will provide us to determine the coordinates of the point L_d .

Since $m_{d_{BC}} = \frac{y_0 - b}{a - x_0}$, by using the fact of the product of the tangents of two perpendicular lines, it can be obtained that $m_{d_{LL_d}} = -\frac{(a - x_0)}{y_0 - b}$.

Both of the lines' equations can be constructed by using a point which belongs to the line and its tangent.

The equation of the line d_{BC} can be calculated as follows;

$$d_{BC} : y - y_0 = \frac{y_0 - b}{a - x_0}(x - a)$$

$$\Rightarrow (y - y_0)(a - x_0) = (y_0 - b)(x - a)$$

After the revision of equation, following form can be obtained;

$$y(a - x_0) + x(b - y_0) = a(b - y_0) + y_0(a - x_0) \quad (7)$$

The equation of the line d_{LL_d} can be calculated as follows;

$$d_{LL_d} : y - b = -\frac{a - x_0}{y_0 - b}(x - a)$$

$$\Rightarrow (y - b)(y_0 - b) = -(a - x_0)(x - a)$$

After the revision of equation, following form can be obtained;

$$y(y_0 - b) - (x_0 - a)x = a(a - x_0) + b(y_0 - b) \quad (8)$$

By finding the common solution of the equation (7) and (8) the coordinates of L_d can be obtained as follows;

$$L_d \left(\frac{a[(y_0 - b)^2 + (a - x_0)^2] - (a - x_0)(y_0 - b)^2}{(a - x_0)^2 + (y_0 - b)^2}, \frac{y_0(a - x_0)^2 + b(b - y_0)^2}{(a - x_0)^2 + (b - y_0)^2} \right) \quad (9)$$

Finding the coordinates of the point M_d :

The intersection point of the line d_{DC} and the line d_{MM_d} will provide us to determine the coordinates of the point M_d .

Since $m_{d_{DC}} = \frac{b - y_0}{x_0}$, by using the fact of the product of the tangents of two perpendicular lines, it can be obtained that $m_{d_{MM_d}} = -\frac{x_0}{b - y_0}$.

Both of the lines' equations can be constructed by using a point which belongs to the line and its tangent.

The equation of the line d_{DC} can be calculated as follows;

$$d_{DC} : y - y_0 = \frac{b - y_0}{x_0}x \Rightarrow yx_0 - y_0x_0 = bx - xy_0.$$

After the revision of equation, following form can be obtained;

$$yx_0 + x(y_0 - b) = x_0y_0 \quad (10)$$

The equation of the line d_{MM_d} can be calculated as follows;

$$d_{MM_d} : y - b = -\frac{x_0}{b - y_0}x \Rightarrow (y - b)(b - y_0) = -x_0x$$

$$\Rightarrow yb - yy_0 - b^2 + by_0 = -x_0x$$

After the revision of equation, following form can be obtained;

$$y(b - y_0) + xx_0 = b(b - y_0) \quad (11)$$

By finding the common solution of the equation (10) and (11) the coordinates of M_d can be obtained as follows

$$M_d \left(\frac{x_0(b - y_0)^2}{x_0^2 + (b - y_0)^2}, \frac{x_0^2 y_0 + b(b - y_0)^2}{x_0^2 + (b - y_0)^2} \right) \quad (12)$$

When the triangles DOA, AKB, BLC and CMD are folded perpendicular to the floor on the sides DA, AB, BC and DC respectively, the corner points O, K, L and M will be in their new places. Let's call these revised points as O_t, K_t, L_t and M_t respectively. These points will be in three dimensional space and the coordinates of O_d, K_d, L_d and M_d will be the first two components of them. The third components of each point will be the lengths of $|OO_d|, |KK_d|, |LL_d|$ and $|MM_d|$ (figure-2)

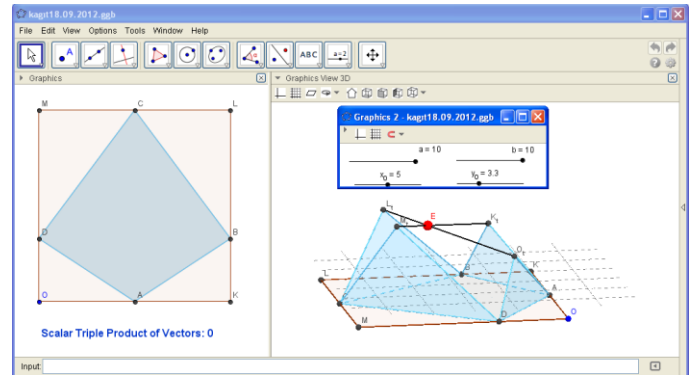


Figure 4: Comparing two and three dimensional position of folded paper.

These lengths can be calculated by using the points O, K, L and M and the equations of the lines d_{DA}, d_{AB}, d_{BC} and d_{DC} . After the proper calculations the lengths are found as follows;

$$|OO_d| = \frac{|x_0 y_0|}{\sqrt{y_0^2 + x_0^2}} \quad |KK_d| = \frac{|(a - x_0) y_0|}{\sqrt{(a - x_0)^2 + y_0^2}}$$

$$|LL_d| = \frac{|(a - x_0)(b - y_0)|}{\sqrt{(a - x_0)^2 + (b - y_0)^2}} \quad |MM_d| = \frac{|x_0(b - y_0)|}{\sqrt{x_0^2 + (b - y_0)^2}}$$

So, the ordered triples as the components of the points O_t, K_t, L_t and M_t can be stated as follows;

$$O_t \left(\frac{x_0 y_0^2}{x_0^2 + y_0^2}, \frac{x_0^2 y_0}{x_0^2 + y_0^2}, \frac{|x_0 y_0|}{\sqrt{y_0^2 + x_0^2}} \right)$$

$$K_i \left(\frac{y_0^2 x_0 + a(a-x_0)^2}{y_0^2 + (a-x_0)^2}, \frac{y_0(a-x_0)^2}{y_0^2 + (a-x_0)^2}, \frac{|(a-x_0)y_0|}{\sqrt{(a-x_0)^2 + y_0^2}} \right)$$

The components of the $L_i = (L_{ix}, L_{iy}, L_{iz})$

$$L_{ix} = \frac{a[(y_0-b)^2 + (a-x_0)^2] - (a-x_0)(y_0-b)^2}{(a-x_0)^2 + (y_0-b)^2}$$

$$L_{iy} = \frac{y_0(a-x_0)^2 + b(b-y_0)^2}{(a-x_0)^2 + (b-y_0)^2}$$

$$L_{iz} = \frac{|(a-x_0)(b-y_0)|}{\sqrt{(a-x_0)^2 + (b-y_0)^2}}$$

$$M_i \left(\frac{x_0(b-y_0)^2}{x_0^2 + (b-y_0)^2}, \frac{x_0^2 y_0 + b(b-y_0)^2}{x_0^2 + (b-y_0)^2}, \frac{|x_0(b-y_0)|}{\sqrt{x_0^2 + (b-y_0)^2}} \right)$$

The components of the vector $O_i L_i$;

$$\left(\frac{a[(y_0-b)^2 + (a-x_0)^2] - (a-x_0)(y_0-b)^2}{(a-x_0)^2 + (y_0-b)^2} - \frac{x_0 y_0^2}{x_0^2 + y_0^2}, \frac{y_0(a-x_0)^2 + b(b-y_0)^2}{(a-x_0)^2 + (b-y_0)^2} - \frac{x_0^2 y_0}{x_0^2 + y_0^2}, \frac{|(a-x_0)(b-y_0)|}{\sqrt{(a-x_0)^2 + (b-y_0)^2}} - \frac{|x_0 y_0|}{\sqrt{y_0^2 + x_0^2}} \right)$$

The components of the vector $K_i M_i$;

$$\left(\frac{x_0(b-y_0)^2}{x_0^2 + (b-y_0)^2} - \frac{y_0^2 x_0 + a(a-x_0)^2}{y_0^2 + (a-x_0)^2}, \frac{x_0^2 y_0 + b(b-y_0)^2}{x_0^2 + (b-y_0)^2} - \frac{y_0(a-x_0)^2}{y_0^2 + (a-x_0)^2}, \frac{|x_0(b-y_0)|}{\sqrt{x_0^2 + (b-y_0)^2}} - \frac{|(a-x_0)y_0|}{\sqrt{(a-x_0)^2 + y_0^2}} \right)$$

The components of the vector $O_i K_i$;

$$\left(\frac{y_0^2 x_0 + a(a-x_0)^2}{y_0^2 + (a-x_0)^2} - \frac{x_0 y_0^2}{x_0^2 + y_0^2}, \frac{y_0(a-x_0)^2}{y_0^2 + (a-x_0)^2} - \frac{x_0^2 y_0}{x_0^2 + y_0^2}, \frac{|(a-x_0)y_0|}{\sqrt{(a-x_0)^2 + y_0^2}} - \frac{|x_0 y_0|}{\sqrt{y_0^2 + x_0^2}} \right)$$

The triple product of the vectors can be calculated by following determinant;

$$\begin{vmatrix} \frac{a[(y_0-b)^2 + (a-x_0)^2] - (a-x_0)(y_0-b)^2}{(a-x_0)^2 + (y_0-b)^2} - \frac{x_0 y_0^2}{x_0^2 + y_0^2} & \frac{y_0(a-x_0)^2 + b(b-y_0)^2}{(a-x_0)^2 + (b-y_0)^2} - \frac{x_0^2 y_0}{x_0^2 + y_0^2} & \frac{|(a-x_0)(b-y_0)|}{\sqrt{(a-x_0)^2 + (b-y_0)^2}} - \frac{|x_0 y_0|}{\sqrt{y_0^2 + x_0^2}} \\ \frac{x_0(b-y_0)^2}{x_0^2 + (b-y_0)^2} - \frac{y_0^2 x_0 + a(a-x_0)^2}{y_0^2 + (a-x_0)^2} & \frac{x_0^2 y_0 + b(b-y_0)^2}{x_0^2 + (b-y_0)^2} - \frac{y_0(a-x_0)^2}{y_0^2 + (a-x_0)^2} & \frac{|x_0(b-y_0)|}{\sqrt{x_0^2 + (b-y_0)^2}} - \frac{|(a-x_0)y_0|}{\sqrt{(a-x_0)^2 + y_0^2}} \\ \frac{y_0^2 x_0 + a(a-x_0)^2}{y_0^2 + (a-x_0)^2} - \frac{x_0 y_0^2}{x_0^2 + y_0^2} & \frac{y_0(a-x_0)^2}{y_0^2 + (a-x_0)^2} - \frac{x_0^2 y_0}{x_0^2 + y_0^2} & \frac{|(a-x_0)y_0|}{\sqrt{(a-x_0)^2 + y_0^2}} - \frac{|x_0 y_0|}{\sqrt{y_0^2 + x_0^2}} \end{vmatrix} = 0$$

IV. CONCLUSION

In the GeoGebra 5.0 Beta release platform, it has been easily seen that the scalar triple product is zero when the intersected position of the vectors $O_i L_i$ and $K_i M_i$ is captured. By this way, students may understand the relationship of their fore known mathematical knowledge and a real life situation [6].

Now, we have the points O_i, K_i, L_i and M_i in space. We can think as the mathematical question in our main problem is “what is the condition for the line segments $O_i L_i$ and $K_i M_i$ are intersected?”. The basic answer will be the vectors $O_i L_i$ and $K_i M_i$ must be coplanar and not parallel. Since the position of the points O_i, K_i, L_i ve M_i in space, we can be sure that these vectors are not parallel and has intersection points when they are reflected on the floor. Let’s check the vectors $O_i L_i$ and $K_i M_i$ coplanar when they are intersected in space.

For this check, the fact of “three vectors’ scalar triple product must be zero to be coplanar” can be used. We have two vectors for this operation. By choosing the vector $O_i K_i$, additionally, we can calculate the scalar triple product.

There is also another opportunity of understanding the relationship between algebraic and geometric representations in this paper [7] although it still needs to be proved. We observed that at least one points of A and B must be the midpoint of the segment which the point belongs. This fact can be proved by solving the proper equations obtained from the scalar triple product.

REFERENCES

- [1] AKKOÇ, H. (2006). Fonksiyon Kavramının Çoklu Temsillerinin Çağrıştırdığı Kavram Görüntüleri. H.Ü. Eğitim Fakültesi Dergisi (H.U. Journal of Education). 30. 1-10
- [2] AKTÜMEN, M. & KABACA, T. (2012). Exploring the Mathematical Model Of The Thumbaround Motion by Geogebra. Technology, Knowledge and Learning. DOI: 10.1007/s10758-012-9194-5
- [3] AKTUMEN, M., BALTACI, S. & YILDIZ, A. (2011). Calculating the surface area of the water in a rolling cylinder and visualization as two and three dimensional by means of GeoGebra. International Journal of Computer Applications (www.ijcaonline.org/archives/volume25/number1/3170-4022)
- [4] KABACA, T. & AKTUMEN, M. (2010), Using Geogebra as an Expressive Modeling Tool: Discovering the Anatomy of the Cycloids Parametric Equation, International Journal of GeoGebra The New Language For The Third Millennium, Zigotto Printing And Publishing House, Galati-Romania, Vol.1 No.1, 63-81 Issn: 2068-3227.
- [5] AKTÜMEN, M. HORZUM & T., CEYLAN, T.(2010). Önünde Engel BulunanBirKaleminUcununIzininParametrikDenklemininHesaplanması veGeogebraileGörselleştirme.9. MatematikSempozyumuSergiveŞenlikleri.KaradenizTeknikÜniversitesi, 20-22 Ekim 2010 Trabzon.
- [6] ZBIEK, R. M. & CONNER, A. (2006). Beyond motivation: Exploring mathematical modeling as a context for deepening students' understandings of curricular mathematics. Educational Studies in Mathematics, 63(1), 89-112.
- [7] ÖZMANTAR, M., F., BINGÖLBALI, E. & AKKOÇ, H. (2008). Matematiksel kavramyanılgılarıveçözümlerileri.PegemAkademi. Ankara.

Sensitivity Analysis of Fourier Transformation Spectrometer: FTS Against Observation Noise on Retrievals of Carbon Dioxide and Methane

Kohei Arai

Graduate School of Science and Engineering
Saga University
Saga City, Japan

Hiroshi Okumura

Graduate School of Science and Engineering
Saga University
Saga City, Japan

Takuya Fukamachi

Graduate School of Science and Engineering
Saga University
Saga City, Japan

Shuji Kawakami

JAXA, Japan
Tsukuba City, Japan

Hirofumi Ohyama

JAXA, Japan
Tsukuba City, Japan

Abstract—Sensitivity analysis of Fourier Transformation Spectrometer: FTS against observation noise on retrievals of carbon dioxide and methane is conducted. Through experiments with real observed data and additive noise, it is found that the allowable noise on FTS observation data is less than 2.1×10^{-5} if estimation accuracy of total column carbon dioxide and methane is better than 1(%) .

Keywords-FTS; carbon dioxide, methane, sensitivity analysis, error analysis.

I. INTRODUCTION

Greenhouse gases Observing SATellite: GOSAT carries TANSO CAI for clouds and aerosol particles observation of mission instrument and TANSO FTS¹: Fourier Transformation Spectrometer² for carbon dioxide and methane retrieving mission instrument [1]. In order to verify the retrieving accuracy of two mission instruments, ground based laser radar and TANSO FTS are installed. The former is for TANSO CAI and the latter is for FTS, respectively. One of the other purposes of the ground-based laser radar and the ground-based FTS is to check sensor specifications for the future mission of instruments to be onboard future satellite with extended mission. Although the estimation methods for carbon dioxide and methane are well discussed [2]-[6], estimation method which takes into account measurement noise is not analyzed

yet. Therefore, error analysis for additive noise on estimation accuracy is conducted.

In order to clarify requirement of observation noises to be added on the ground-based FTS observation data, Sensitivity analysis of the ground-based FTS against observation noise on retrievals of carbon dioxide and methane is conducted. Experiments are carried out with additive noise on the real acquired data of the ground-based FTS. Through retrievals of total column of carbon dioxide and methane with the noise added the ground-based FTS signals, retrieval accuracy is evaluated. Then an allowable noise on the ground-based FTS which achieves the required retrieval accuracy (1%) is reduced.

The following section describes the proposed sensitivity analysis followed by some experiments. Then concluding remarks with some discussions is followed by.

II. PROPOSED SENSITIVITY ANALYSIS

A. Ground-based FTS

Figure 1 shows schematic configuration of the ground-based FTS which is originated from Michelson Interference Measurement Instrument. Light from the light source divided in to two directions, the left and the forward at the dichotic mirror of half mirror. The left light is reflected at the fixed hold mirror and reaches to the half mirror while the forward light is reflected at the moving mirror and reaches at the half mirror. Then interference occurs between the left and the forward lights. After that interference light is detected by detector. Outlook of the ground-based FTS is shown in Figure 2.

¹ http://www.jaxa.jp/projects/sat/gosat/index_j.html

² <http://ja.wikipedia.org/wiki/%E3%83%9E%E3%82%A4%E3%82%B1%E3%83%AB%E3%82%BD%E3%83%B3%E5%B9%B2%E6%B8%89%E8%A8%88>

Figure 3 (a) shows an example of the interferogram³ (interference light detected by the detector of the ground-based FTS). By applying Fourier Transformation to the interferogram, observed Fourier spectrum is calculated as shown in Figure 3 (b). When the ground-based FTS observes the atmosphere, the observed Fourier spectrum includes absorptions due to atmospheric molecules and aerosol particles. By comparing to the spectrum which is derived from the radiative transfer code with atmospheric parameters, atmospheric molecules and aerosol particles are estimated.

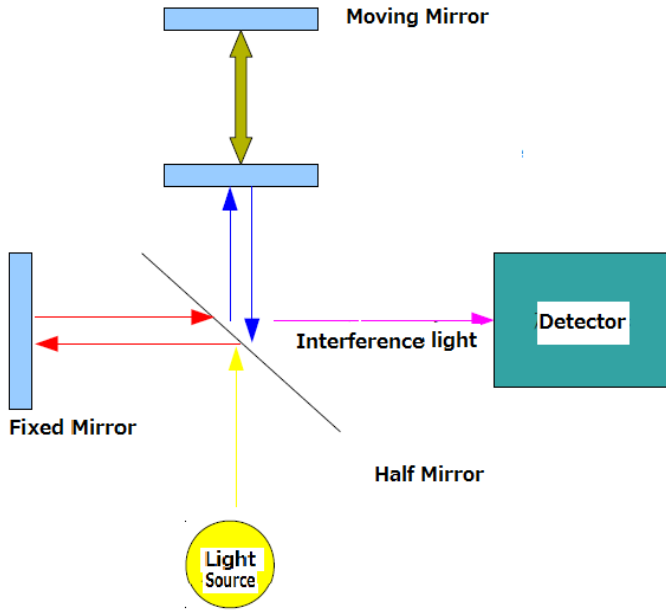


Figure 1 Michelson Interference Measurement Instrument

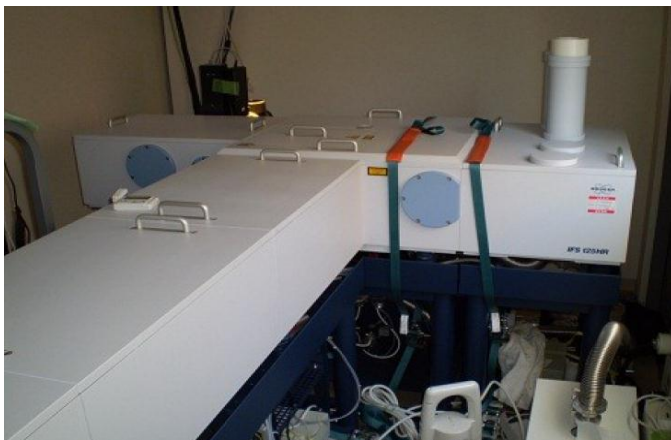
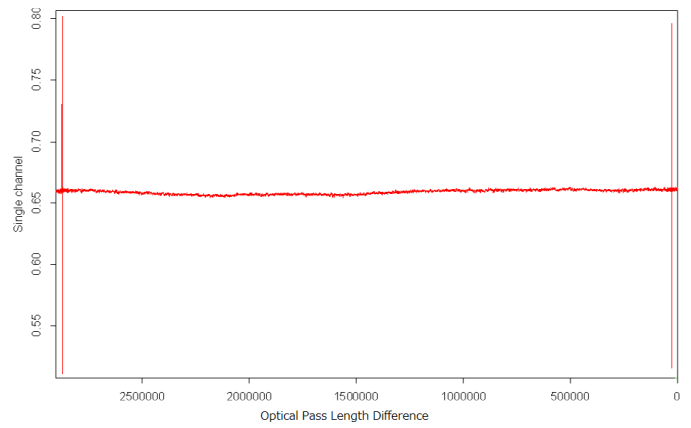
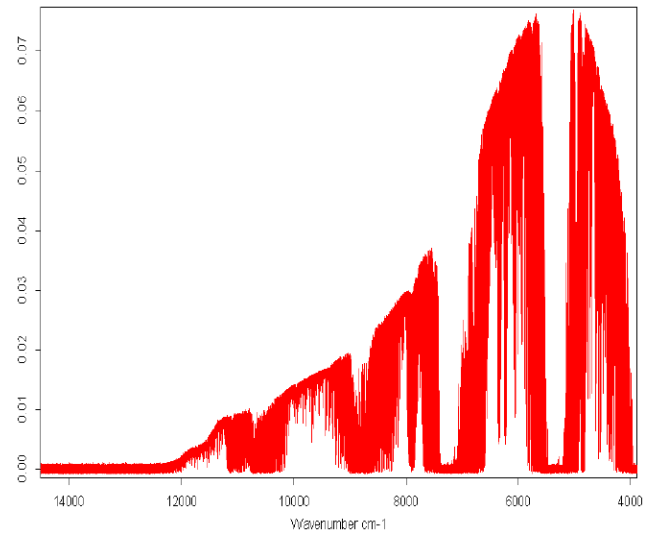


Figure 2 Outlook of the FTS used



(a) Interferogram



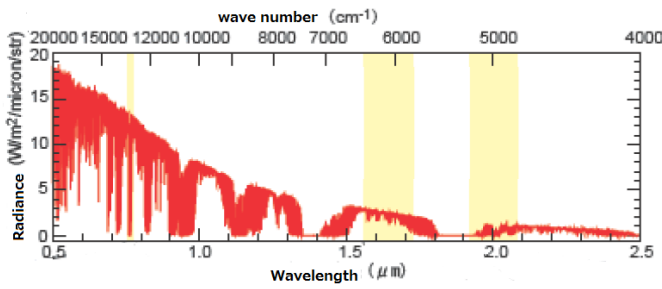
(b) Fourier spectrum

Figure 3 Examples of interferogram and Fourier spectrum when FTS observes the atmosphere

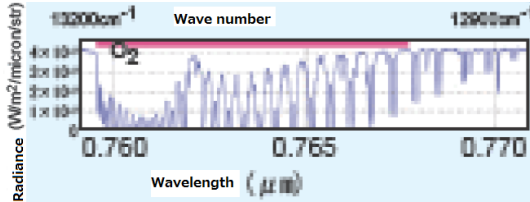
B. Principle for Carbon Dioxide and Methane Retrievals with TANSO FTS Data

Figure 4 shows a principle of the retrieval method for atmospheric constituents using GOSAT/TANSO data. Figure 4 (a) shows Top of the Atmosphere: TOA radiance in the wavelength ranges from 500 to 2500nm (visible to shortwave infrared wavelength regions). There are three major absorption bands due to oxygen (760-770nm), carbon dioxide and methane (1600-1700nm), and water vapor and carbon dioxide (1950-2050nm) as shown in Figure 4 (b), (c), and (d), respectively. These bands are GOSAT/TANSO spectral bands, Band 1 to 3, respectively. In addition to these, there is another wide spectrum of spectral band, Band 4 as shown in Figure 4(e) which covers from visible to thermal infrared regions..

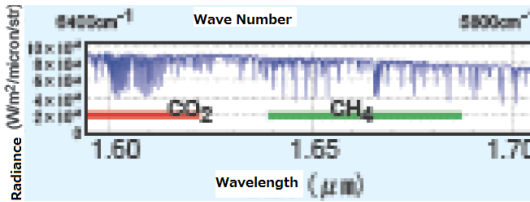
³<http://en.wikipedia.org/wiki/Interferometry>



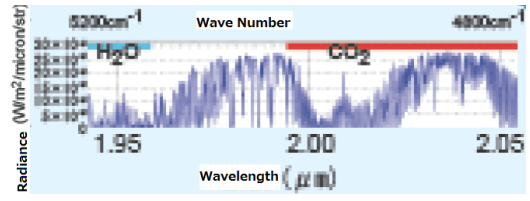
(a) TOA radiance



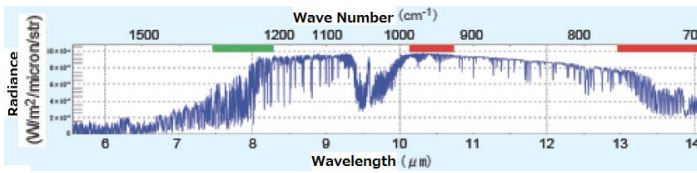
(b) Band 1



(c) Band 2



(d) Band 3



(e) Band 4

Figure 4 Example of TOA radiance and absorption bands as well as spectral bands of GOSAT/TANSO instrument

III. EXPERIMENTS

A. Ground-based FTS Data Used

The ground-based FTS data used for experiments are acquired on November 14 and December 19 2011. Figure 5 shows the interferograms derived from the acquired the ground-based FTS data.

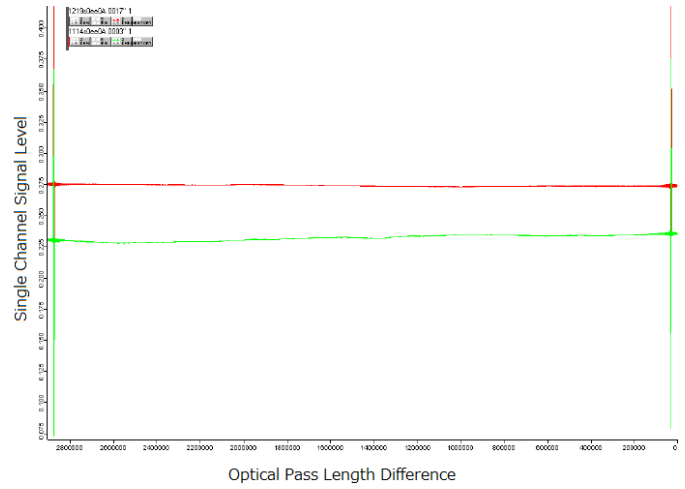


Figure 5 Example of interferograms used for experiments.

B. Experimental Method

Observation noise is included in the observed interferograms. In addition to the existing noise, several levels of additional noises which are generated by random number generator of Messene Twister with zero mean and several standard deviations is added on to the iterferograms as shown in Figure 6.

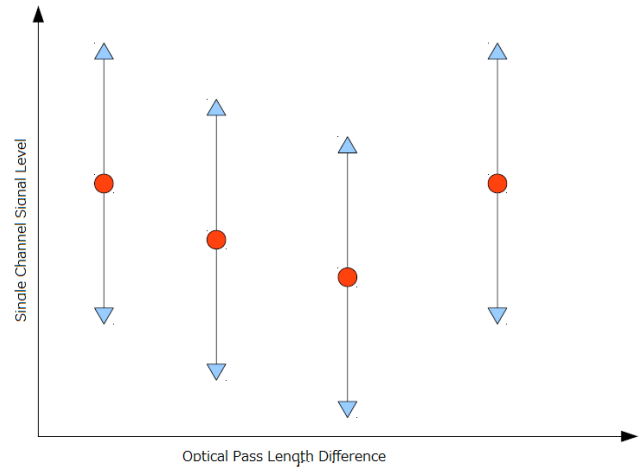
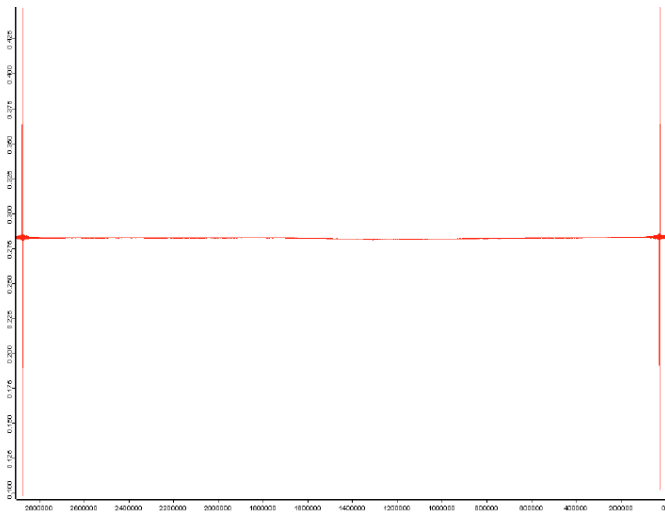


Figure 6 Method for adding the noises to the acquired interferograms

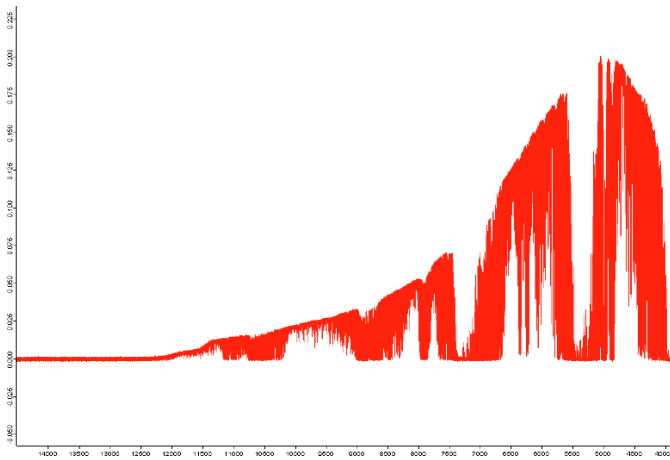
C. Experimental Results (Noise Added Interferograms and Fourier Spectra)

Figure 7 shows noise added interferograms and the Fourier spectra derived from the noise added interferograms. Added noises ranges from 0 to 1×10^{-3} . $1/1000$ of standard deviation of noise (zero mean) against signal level is added to the single channel of signal level in maximum. Vertical axis shows signal level and horizontal axis shows optical pass length difference for Figure 7 (a), (c), (e), (g), (i), (k), and (m) while vertical axis shows Fourier spectrum (amplitude) and horizontal axis shows frequency (or wave number) for the rest of Figure 7.

As shown in Figure 7, Fourier spectra is degrading in accordance with increasing of noise obviously. Although the additive noises are not clearly seen, it is slightly recognizable the noise through comparison between Figure (b) and (c).



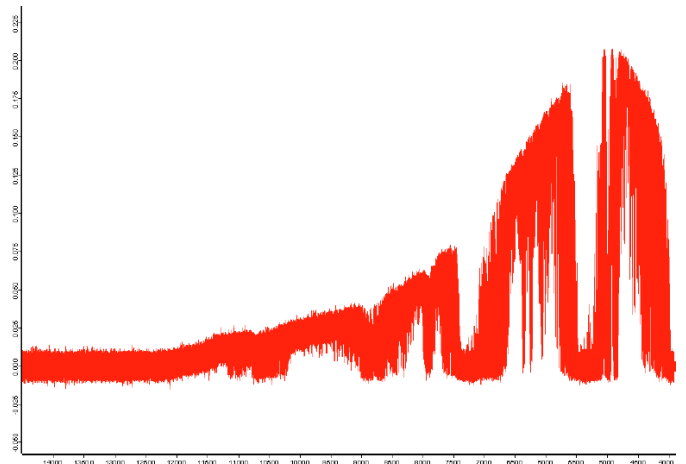
(a)Interferogram(0)



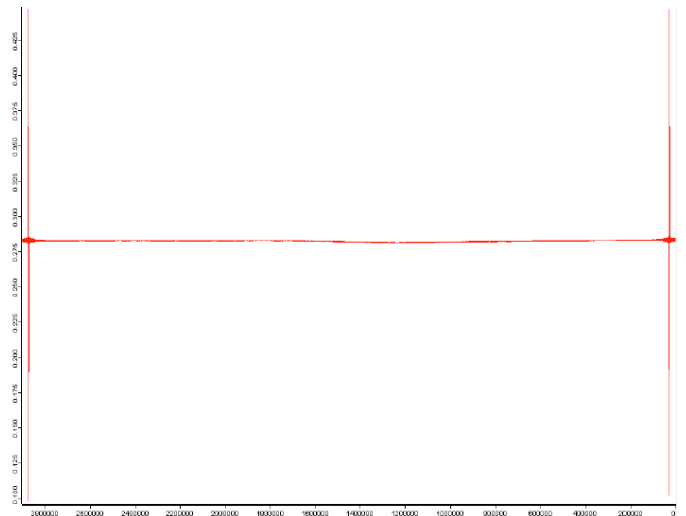
(b)Fourier spectrum(0)



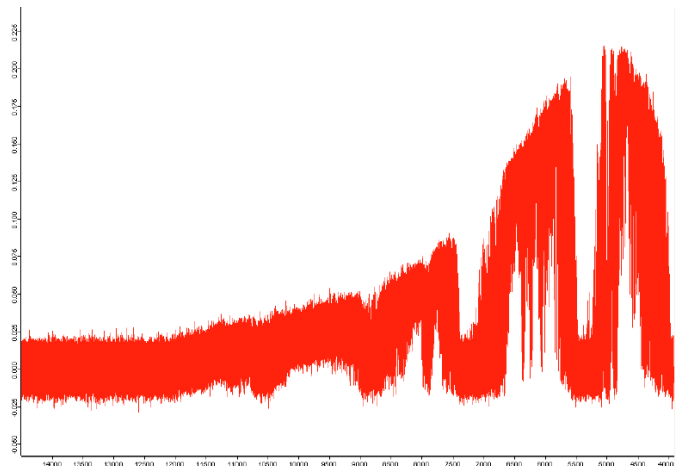
(c)Interferogram(2.5×10^{-5})



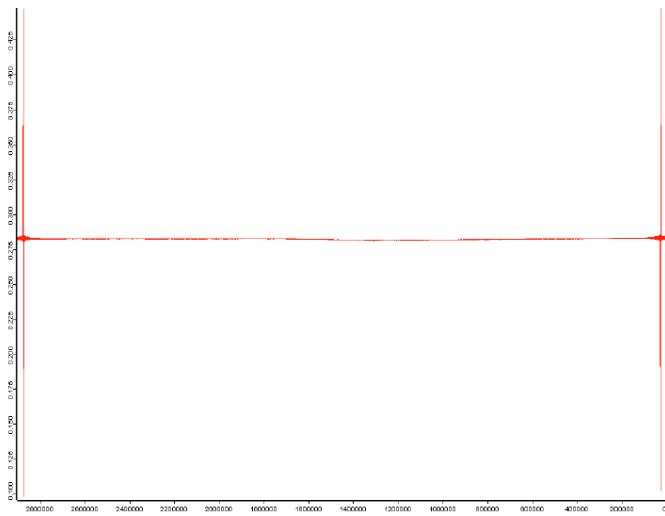
(d)Fourier spectrum(2.5×10^{-5})



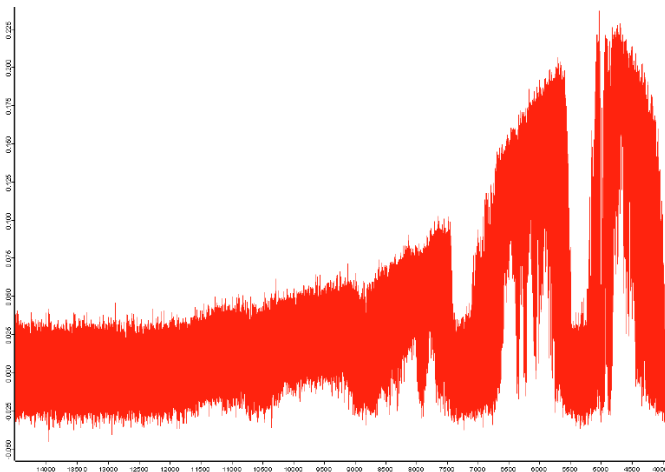
(e)Interferogram(5×10^{-5})



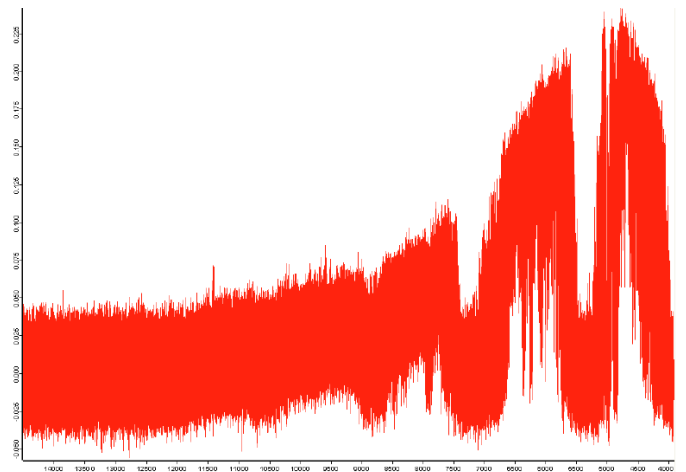
(f)Fourier spectrum(5×10^{-5})



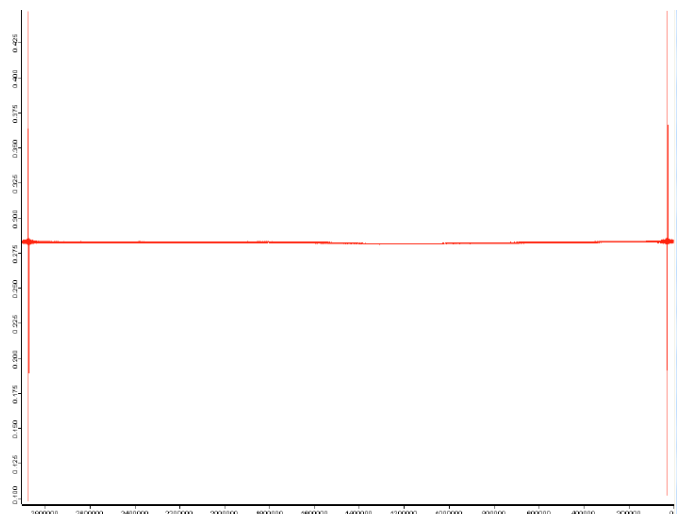
(g) Interferogram(7.5×10^{-5})



(h) Fourier spectrum(7.5×10^{-5})



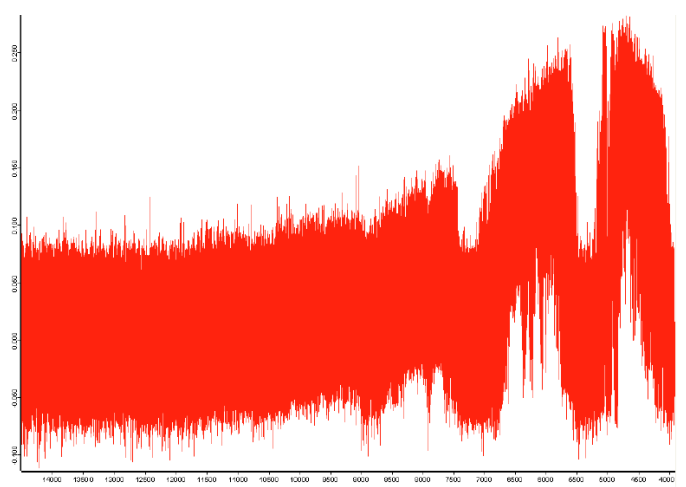
(j) Fourier spectrum(1×10^{-4})



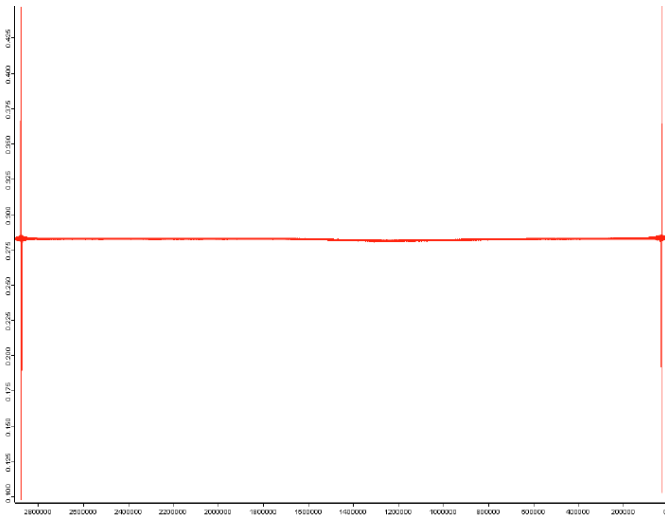
(k) Interferogram(2×10^{-4})



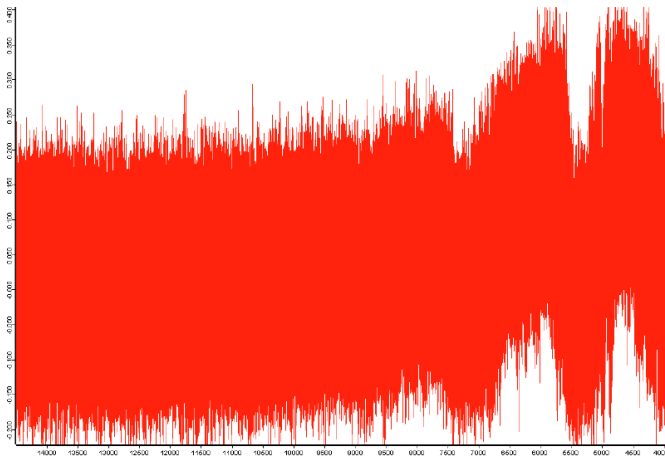
(i) Interferogram(1×10^{-4})



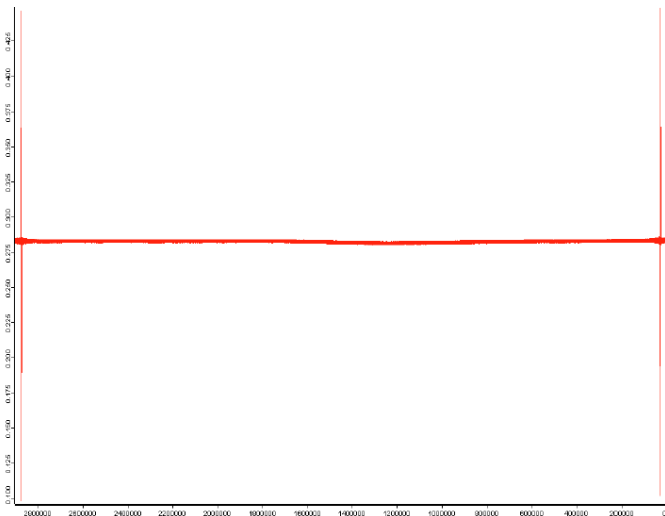
(l) Fourier spectrum(2×10^{-4})



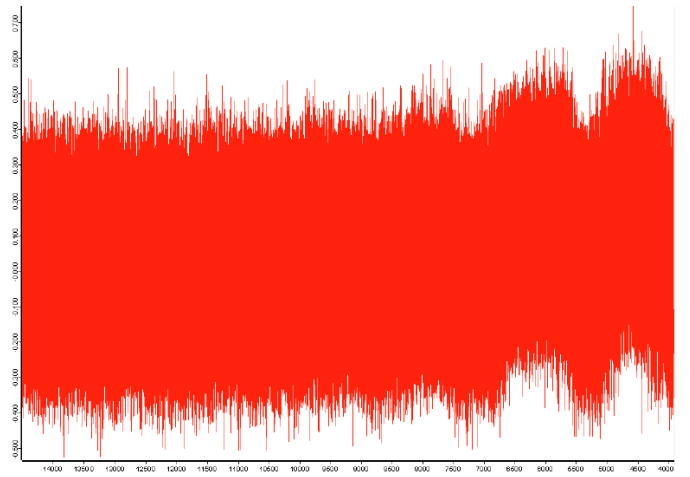
(m) Interferogram(5×10^{-4})



(n) Fourier spectrum(5×10^{-4})



(o) Interferogram(1×10^{-3})



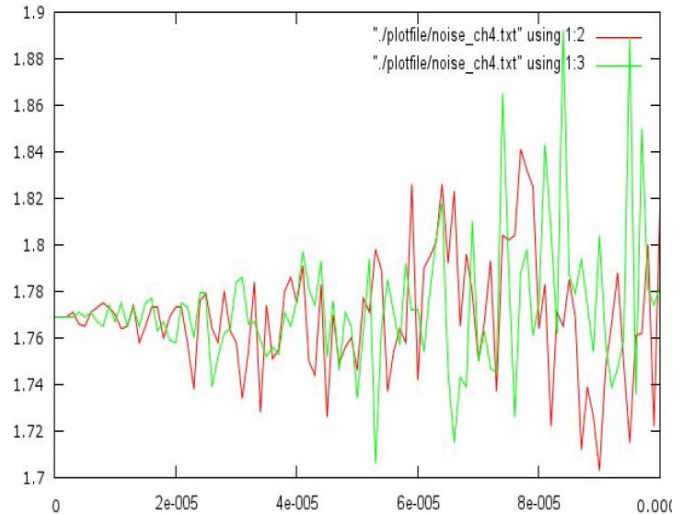
(p) Fourier spectrum(1×10^{-3})

Figure 7 Noise added Interferograms and Fourier spectrum derived from the interferograms

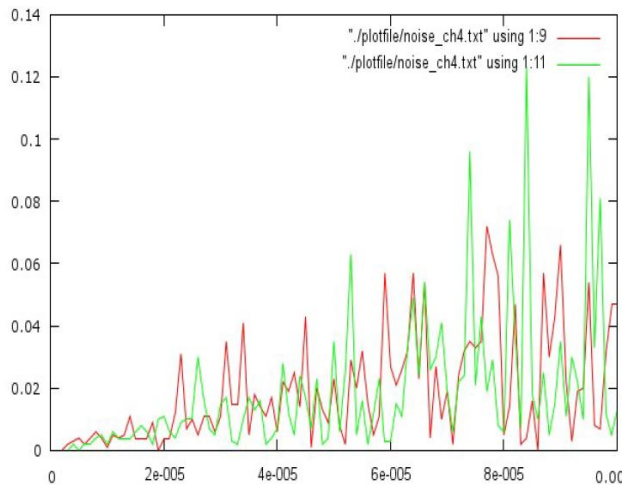
D. Experimental Results (Retrieval Error)

Figure 8 (a) shows methane retrieved results. Horizontal axis shows standard deviation of additive noise and vertical axis shows retrieved methane amount in unit of ppm (percent per million). Figure 8 (b) shows retrieved error (retrieved methane amount from noise added interferogram minus retrieved methane amount from noise free interferogram).

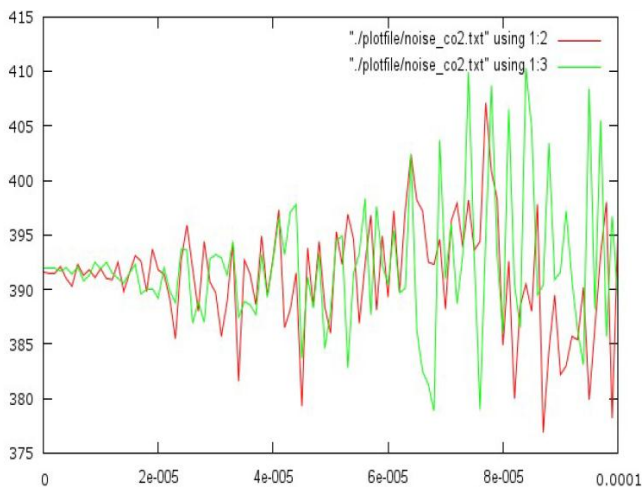
Meanwhile, Figure 8 (c) shows carbon dioxide retrieved results. Horizontal axis shows standard deviation of additive noise and vertical axis shows retrieved carbon dioxide amount in unit of ppm (percent per million). Figure 8 (d) shows retrieved error (retrieved methane amount from noise added interferogram minus retrieved carbon dioxide amount from noise free interferogram). GFIT of retrieval software code is used for both estimations of total column carbon dioxide and methane contents in the atmosphere [6].



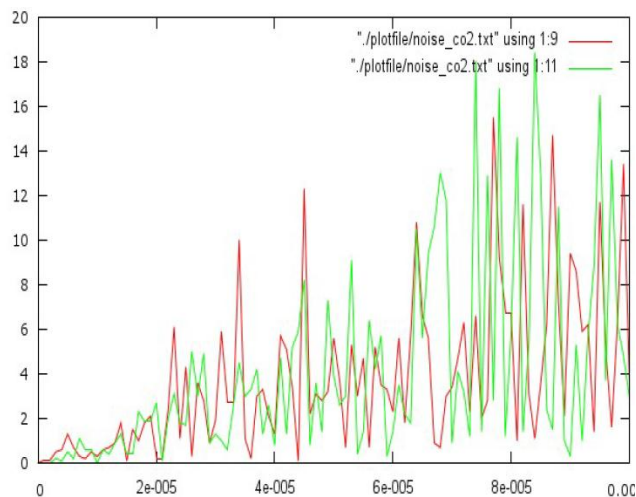
(a) Retrieved methane amount



(b) Retrieved error



(c) Retrieved carbon dioxide amount



(d) Retrieved error

Figure 8 Retrieved carbon dioxide and methane amount with noise added and noise free iterferograms together with retrieved errors.

From these figures, it is concluded as follows,

Allowable retrieval errors for methane and carbon dioxide are 0.02 ppm and 4 ppm, respectively. Therefore, acceptable noise level on FTS interferogram is less than 2.1×10^{-5} .

IV. CONCLUSION

Sensitivity analysis of Fourier Transformation Spectrometer: FTS against observation noise on retrievals of carbon dioxide and methane is conducted. Through experiments with real observed data and additive noise, it is found that the allowable noise on FTS observation data is less than 2.1×10^{-5} if estimation accuracy of total column carbon dioxide and methane is better than 1% (allowable retrieval errors for methane and carbon dioxide are 0.02 ppm and 4 ppm, respectively. These correspond to 1% error for both methane and carbon dioxide retrievals).

REFERENCES

- [1] http://repository.tksc.jaxa.jp/dr/prc/japan/contents/AA0065136000/65136000.pdf?IS_STYLE=jpn (Accessed on September 14 2012)
- [2] Clough, S. A., et al. [2006], IEEE Trans. Geosci. Remote Sens., 44, 1308–1323.
- [3] Hase, F., et al. [2004], J. Quant. Spectrosc. Radiat. Transfer, 87(1), 25–52.
- [4] Rodgers, C. D. [2000], Inverse Methods for Atmospheric Sounding: Theory and Practice.
- [5] Tikhonov, A. [1963], Dokl. Acad. Nauk SSSR, 151, 501–504. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [6] Wunch, D, et al., [2001], The Total Carbon Cloumn Obsrving Network (TCCON), Phil., Tarns. R. Soc. A., 369, 2087–2112.

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science, and Technology of the University of Tokyo from 1974 to 1978 also was with National Space Development Agency of Japan (current JAXA) from 1979 to 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He was appointed professor at Department of Information Science, Saga University in 1990. He was appointed councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was also appointed councilor of Saga University from 2002 and 2003 followed by an executive councilor of the Remote Sensing Society of Japan for 2003 to 2005. He is an adjunct professor of University of Arizona, USA since 1998. He also was appointed vice chairman of the Commission “A” of ICSU/COSPAR in 2008. He wrote 30 books and published 332 journal papers.

Sensitivity Analysis for Water Vapor Profile Estimation with Infrared: IR Sounder Data Based on Inversion

Kohei Arai¹

Graduate School of Science and Engineering
Saga University
Saga City, Japan

Abstract— Sensitivity analysis for water vapor profile estimation with Infrared: IR sounder data based on inversion is carried out. Through simulation study, it is found that influence due to ground surface relative humidity estimation error is greater than that of sea surface temperature estimation error on the water vapor vertical profile retrievals.

Keywords- IR sounder; error budget analysis; MODTRAN; air temperature; relative humidity.

I. INTRODUCTION

Air-temperature and water vapor profiles are used to be estimated with Infrared Sounder data [1]. One of the problems on retrieving vertical profiles is its retrieving accuracy. In particular, estimation accuracy of air-temperature and water vapor at tropopause¹ altitude is not good enough because there are gradient changes of air-temperature and water vapor profile in the tropopause so that observed radiance at the specific channels are not changed for the altitude.

In order to estimate air-temperature and water vapor, least square based method is typically used. In the process, Root Mean Square: RMS difference between observed radiance and calculated radiance with the designated physical parameters are minimized. Then the designated physical parameters including air-temperature and water vapor at the minimum RMS difference are solutions.

Typically, Newton-Raphson method² which gives one of local minima is used for minimization of RMS difference. Newton-Raphson needs first and second order derivatives, Jacobean and Hessian at around the current solution. It is not easy to formularize these derivatives analytically. The proposed method is based on Levenberg Marquardt: LM of non-linear least square method³. It uses numerically calculated first and second order derivatives instead of analytical based derivatives. Namely, these derivatives can be calculated with radiative transfer model based radiance calculations. At around the current solution in the solution space, directional derivatives are calculated with the radiative transfer model.

The proposed method is validated for air-temperature and water vapor profile retrievals with Infrared: IR sounder⁴ data derived from AQUA/AIRS [2]-[7]. A comparison of retrieving accuracy between Newton-Raphson method and the proposed method based on LM method [8] is made in order to demonstrate an effectiveness of the proposed method in terms of estimation accuracy in particular for the altitude of tropopause [9]. Global Data Assimilation System: GDAS⁵ data of assimilation model derived 1 degree mesh data is used as truth data of air-temperature and water vapor profiles. The experimental data show that the proposed method is superior to the conventional Newton-Raphson method.

Atmospheric sounding can be improved by using the high spectral resolution of sounder, such as AIRS, IASI, instead of HIRS, TOVS used in the past three and half decades[10]-[13]. Their sensors have large number of channels, and have large amount of atmospheric sounding information in the measurement data. But for the retrieval of air temperature profile, it is not practical nor an advantage to use all spectral points. Therefore it is important for this work to eliminate those channels whose information does not add to the final retrieval accuracy and even before for the sake of efficiency, those channels potentially contaminated by solar radiation or significantly affected by other gases (not required for temperature profiling).

Sensitivity analysis of water vapor profile retrieval against surface relative humidity and sea surface temperature is carried out. Multi channels utilized retrieval method is assumed for water vapor profile estimation. Then influences due to surface relative humidity and sea surface temperature on water vapor profile retrieval accuracy are investigated.

The following section describes background of the research as well as assumptions for sensitivity analysis followed by detailed method for sensitivity analysis. Secondly, experiments are described with simulation results followed by conclusion and some discussions.

¹ <http://en.wikipedia.org/wiki/Tropopause>

² http://en.wikipedia.org/wiki/Newton's_method

³

http://en.wikipedia.org/wiki/Levenberg%E2%80%93Marquardt_algorithm

⁴ http://en.wikipedia.org/wiki/Atmospheric_Infrared_Sounder

⁵ <http://www.mmm.ucar.edu/mm5/mm5v3/data/gdas.html>

II. RESEARCH BACKGROUND AND METHIOD FOR SENSITIVITY ANALYSIS

A. Infrared: IR Sounder

Sounder instruments on Earth observation satellite are for ozone, water vapor and air temperature vertical profile retrievals with very narrow bandwidth of wavelength channels which measure radiation from the Earth. There are many absorption bands due to atmospheric molecule in infrared and microwave wavelength regions. Atmospheric optical depths are varied with altitude above sea level. Therefore, vertical profile of density of the atmospheric molecules can be estimated results in ozone, water vapor, air temperature profiles can be retrieved.

Table 1 shows major purposes, absorption molecules, and the most appropriate wavelength or frequency regions.

TABLE I. MAJOR PURPOSES, ABSORPTION MOLECULES, AND THE MOST APPROPRIATE WAVELENGTH OR FREQUENCY REGIONS FOR SOUNDER

Wavelength_Region	Wavelength_Frequency	Absorption_Molecule	Major_Purpose
Infrared	6.3um	H2O	Water_vapor
Infrared	9.6um	O3	Ozone
Infrared	15um	CO2	Air Temperature
Microwave	22.235GHz	H2O	Water_vapor
Microwave	60GHz	O2	Air Temperature

B. Profile Retrieval Method

IR sounder data derived radiance, R_{0i} and estimated radiance, R_i based on atmospheric model with the parameters of air-temperature and relative humidity.

$$S = \sum_{i=0}^n [R_i - R_{0i}]^2 \quad (1)$$

Thus the geophysical parameter at when the difference of radiance reaches to the minimum is estimated. Widely used and accurate enough atmospheric software code of MODTRAN⁶ is used in the proposed method. Solution update equation of Newton-Raphson method is expressed by equation (2).

$$X_{n+1} = X_n - H^{-1}J(x_n) \quad (2)$$

where H denotes Hessian matrix⁷ which consists of second order derivatives (residual square error, S which is represented by equation (1) by geophysical parameter, air-temperature and relative humidity). Also J denotes Jacobean⁸ which consists of the first derivative of vectors (S by geophysical parameters, x).

On the other hand, solution update equation is expressed by equation (3).

$$X_{n+1} = X_n + (J^T J + I)^{-1} J^T (R_i - R_{0i}) \quad (3)$$

The first derivative is represented by equation (4) while the second order derivative is expressed by equation (5), respectively.

$$\frac{\partial S}{\partial x_i} = -2 \sum_{k=1}^n [R_k - R_{0k}] \frac{\partial R_{0k}}{\partial x_i} \quad (4)$$

$$\frac{\partial^2 S}{\partial x_i \partial x_j} = 2 \sum_{k=1}^n \left[\frac{\partial R_{0k}}{\partial x_i} * \frac{\partial R_{0k}}{\partial x_j} - [R_k - R_{0k}] \frac{\partial^2 R_{0k}}{\partial x_i \partial x_j} \right] \quad (5)$$

On the other hand, the first and second order derivatives of R with x are expressed by equation (6) and (7), respectively.

$$\frac{\partial R_{0k}}{\partial x_i} \leftarrow \text{MODTRAN} \quad (6)$$

$$\frac{\partial^2 R_{0k}}{\partial x_i \partial x_j} \leftarrow \frac{\partial R_{0k}}{\partial x_i} * \frac{\partial R_{0k}}{\partial x_j} \quad (7)$$

In the proposed method, these derivatives are calculated numerically. 2% changes of relative humidity is taken into account for calculation of derivative R and S while 0.5K changes of air-temperature is also considered for calculation of derivative of R and S . Thus the geophysical parameter at when the difference of radiance reaches to the minimum is estimated.

C. Weighting Function

Based on radiative transfer model, upwelling radiance from the atmosphere is greater than ground surface radiance in accordance with absorption molecule density in the case that the atmosphere contains a plenty of absorption molecules.

Weighting function which represents contributions to upwelling radiance at all the altitudes can be determined by optical depth. Namely,

$$K(z) = \int_0^{\infty} \frac{\partial e^{-\tau(p,\rho)}}{\partial p} dp$$

$$tr(z) = e^{(-\int_z^{\infty} \rho k_y dz)} = e^{-\tau} \quad (8)$$

where K denotes weighting function, tr denotes transparency, τ denotes optical depth, and p denotes atmospheric pressure while z denotes altitude. Also ρ denotes molecule density. For instance, weighting function for 6.3, 7.3, 7.5 um wavelength of H2O absorption channels are shown in Figure 1.

⁶ <http://en.wikipedia.org/wiki/MODTRAN>

⁷ http://en.wikipedia.org/wiki/Hessian_matrix

⁸ http://en.wikipedia.org/wiki/Jacobian_matrix_and_determinant

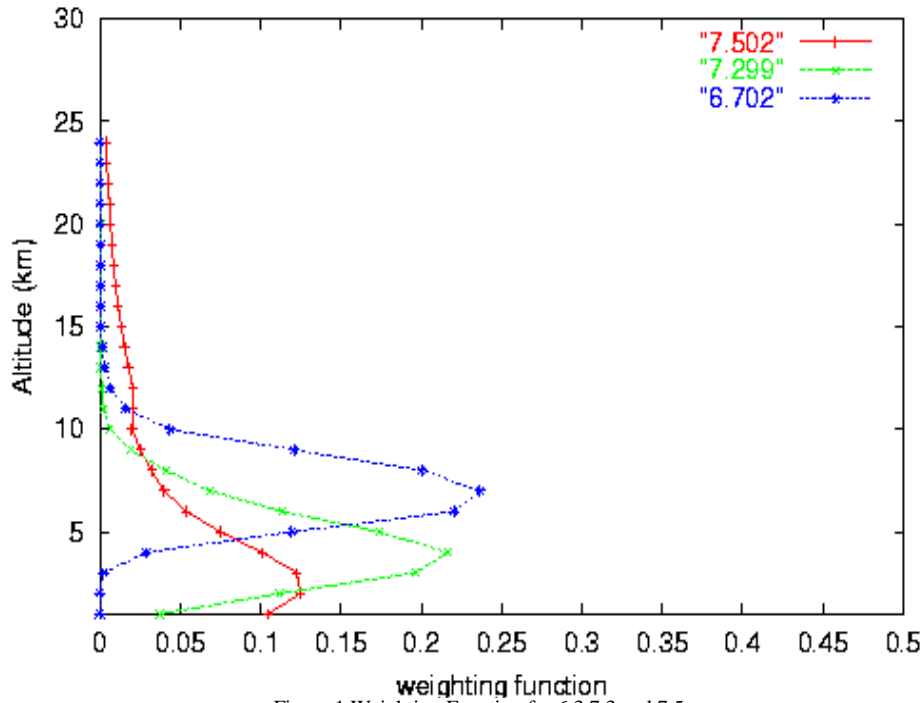


Figure 1 Weighting Function for 6.3,7.3 and 7.5μm

D. Method for Sensitivity Analysis

Assuming MODTRAN is truth, water vapor profile retrieval error is evaluated by which some errors are added to surface air temperature and relative humidity. Then it is possible to clarify relations between water vapor profile retrieval error caused by the errors on the surface air temperature and relative humidity. In the process, steepest descending method of non-linear optimization method for single variable is used for determination of optimum solution of water vapor profile. Equation (9) and (10) shows steepest descending method.

$$(I_{toa} - \hat{I})^2 \leq \varepsilon. \quad (9)$$

$$RH^{n+1} = RH^n + \lambda \frac{dI^n}{dRH^n} \quad (10)$$

Thus the retrieval error can be evaluated.

III. EXPERIMENTS

A. Experimental Parameters

All the parameters required for experiments is listed in the Table 2. Residual errors of water vapor estimation for surface temperature and relative humidity are shown in Figure 2 and 3, respectively. On the other hands, retrieved water vapor profile is shown in Figure 4. ± 5 and ± 10 (%) of errors are added to the default relative humidity in this case.

TABLE II. PARAMETERS FOR EXPERIMENTS

Atmospheric_Model	Mid.Latitude_Summer
Wavelength	6.7,7.3,7.5um
Default_Relative_Humidity	0%, $\pm 10\%$, $\pm 30\%$,
Default_Air_Temperature	0, ± 1 , ± 3 K
Altitude_at_Top_of_Atmosphere	100km

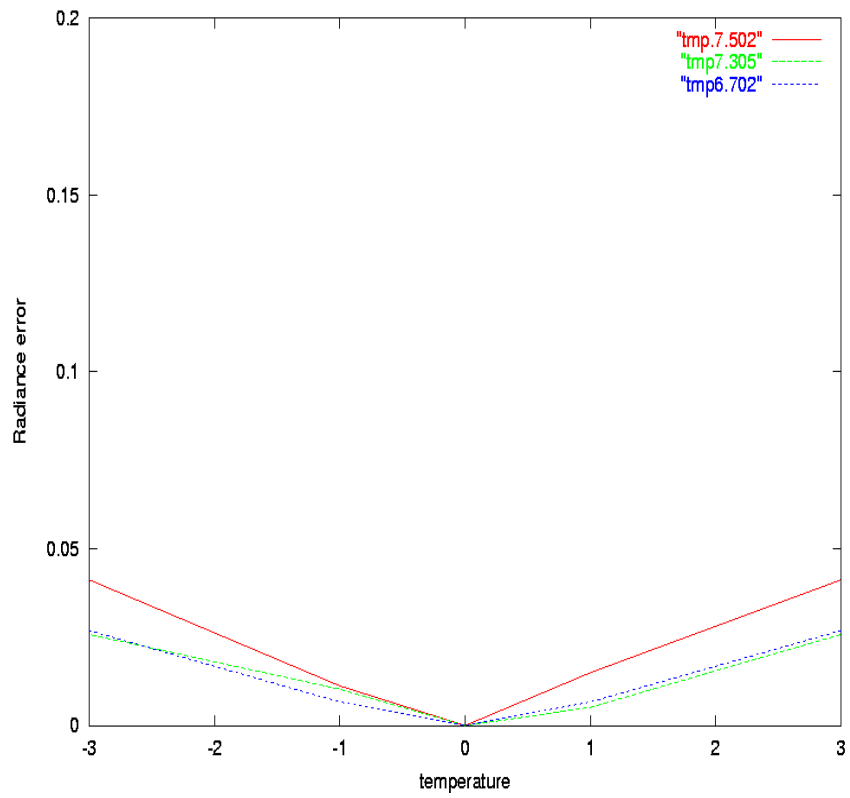


Figure 2 Sensitivity of surface air temperature

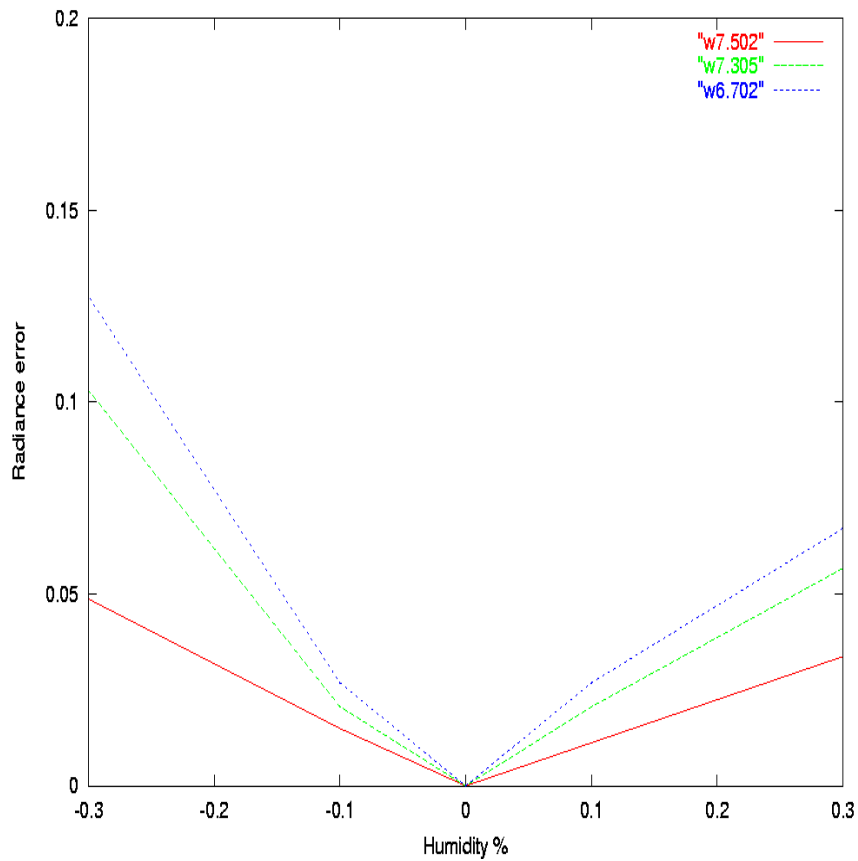


Figure 3 Sensitivity of relative humidity

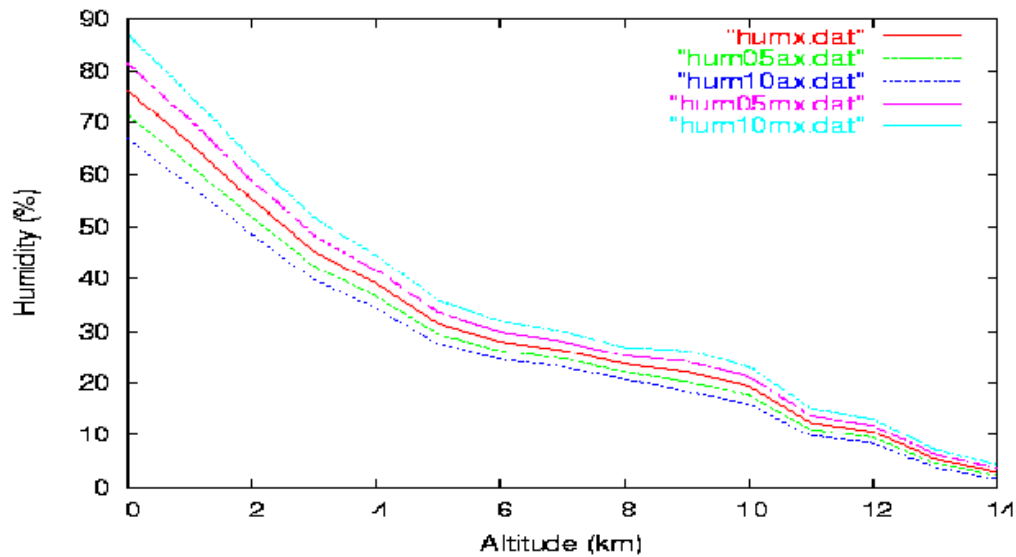


Figure 4 Estimated water vapor profile

B. Summeryzed Experiemntal Results

Through the comparison between Figure 2 and 3, it is found that influence due to error on the relative humidity is greater than that due to surface air temperature. It is reasonable that relative humidity is closely linked to water vapor profile rather than surface air temperature.

Estimated water vapor profile shows also reasonable. In the troposphere, humidity is decreasing smoothly. At around tropopause, there is sharp drop of humidity. This is because that troposphere and stratosphere are interacted weakly each other. Estimated humidity profile shows that $\pm 10\%$ of error which is added to the default humidity is going to be small and greater depending on the altitude. Also it shows a difficulty on water vapor profile retrieval at around tropopause.

IV. CONCLUSION

Sensitivity analysis for water vapor profile estimation with Infrared: IR sounder data based on inversion is carried out. Through simulation study, it is found that influence due to ground surface relative humidity estimation error is greater than that of sea surface temperature estimation error on the water vapor vertical profile retrievals.

It is found that influence due to error on the relative humidity is greater than that due to surface air temperature. It is reasonable that relative humidity is closely linked to water vapor profile rather than surface air temperature.

ACKNOWLEDGMENT

The author would like to thank Mr. Wang King and Dr. Xing Ming Liang, for their effort to experiments.

REFERENCES

[1] Kohei Arai, Lecture Note on Remote Sensing, Morikita-Shuppan publishing Co. Ltd, 2004.

[2] NASA/JPL, "AIRS Overview". NASA/JPL. <http://airs.jpl.nasa.gov/overview/overview/>.

[3] NASA "Aqua and the A-Train". NASA. http://www.nasa.gov/mission_pages/aqua/.

[4] NASA/GSFC "NASA Goddard Earth Sciences Data and Information Services Center". NASA/GSFC. http://disc.gsfc.nasa.gov/AIRS/data_products.shtml.

[5] NASA/JPL "How AIRS Works". NASA/JPL. http://airs.jpl.nasa.gov/technology/how_AIRS_works.

[6] NASA/JPL "NASA/NOAA Announce Major Weather Forecasting Advancement". NASA/JPL. <http://jpl.nasa.gov/news/news.cfm?release=2005-137>.

[7] NASA/JPL "New NASA AIRS Data to Aid Weather, Climate Research". NASA/JPL. <http://www.jpl.nasa.gov/news/features.cfm?feature=1424>.

[8] Kohei Arai and Naohisa Nakamizo, Water vapor and air-temperature profile estimation with AIRS data based on Levenberg-Marquardt, Abstract of the 50th COSPAR(Committee on Space Research/ICSU) Congress, A 3.1-0086-08,995, Montreal, Canada, July, 2008

[9] Kohei Arai and XingMing Liang, sensitivity analysis for air temperature profile estimation method around the tropopause using simulated AQUA/AIRS data, Advances in Space Research, 43, 3, 845-851, 2009.

[10] Jeffrey, A.L., Elisabeth, W., and Gottfried, K., Temperature and humidity retrieval from simulated Infrared Atmospheric Sounding Interferometer(IASI) measurements, J. Geop. Res., 107, 1-11, 2002.

[11] Rodgers, C.D., Information content and optimization of high spectral resolution measurements. In Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research II, vol. 2830, pp. 136-147, Int. Soc. For Optical Eng., Bellingham, Wash.,1996.

[12] Rodgers, C.D., Inverse method for atmospheres: theory and practice, World Sci., Singapore, 2000.

[13] Liou, K.N., An introduction to atmospheric radiation. Elsevier Science, USA, 2002.

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in

Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for

the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission "A" of ICSU/COSPAR since 2008. He wrote 30 books and published 322 journal papers.

Wavelet Based Change Detection for Four Dimensional Assimilation Data in Space and Time Domains

Kohei Arai¹

Graduate School of Science and Engineering
Saga University
Saga City, Japan

Abstract— Method for time change detection of four dimensional assimilation data by means of wavelet analysis is proposed together with spatial change detection with satellite imagery data. The method is validated with assimilation data and the same scale of satellite imagery data. Experimental results show that the proposed method does work well in visually.

Keywords- Assimilation; change detection; wavelet analysis.

I. INTRODUCTION

Data assimilation is useful for numerical prediction of global issues including global warming, weather forecasting, and climate changes [1]-[3]. In particular, four dimensional assimilations, three dimensional space and time dimension, space and time domains, is useful. There are some problems on the current four dimensional assimilations as follows,

- (1) Prediction accuracy is not good enough,
- (2) Precision of input data is not good enough,
- (3) Not so stable solution can be derived
- (4) Boundary conditions cannot be given properly,
- (5) Change detection performance is not good enough.

In order to overcome the aforementioned problems, (3) and (5), wavelet analysis is introduced here in this paper. Namely, wavelet analysis based prediction is attempted for getting stable solutions. Also, wavelet analysis based space and time change detections are attempted as well.

One of the examples is shown in this paper. Time series of three dimensional air temperature and relative humidity are created with assimilation model of which earth observation satellite data derived profiles. These are called as “state variables”. Space and time changes on the state variables are detected in terms of the change locations and change amount. Then extremely hot summer in Japan in 2010 is predicted. It is mainly caused by the jet stream winding in northern hemisphere and the location and magnitude of high pressure system situated in the Pacific Ocean areas. Such this result could be come out from the proposed method.

The following section describes the research background together with the proposed wavelet based method followed by experiments conducted. Then conclusion with some discussions is followed.

II. PROPOSED METHOD

A. Four Dimensional Assimilation

Three dimensional box type of space is assumed for each district in assimilation. In the space, remaining variables, mass, energy and angular momentum are maintained and balanced. Such remaining variable X changed in time domain, $\Delta X/\Delta t$, and is same as supplied variables from the top (F_{top}), from the surface (F_{sfc}), and from the horizontal directions (F_{side}) as shown in Figure 1.

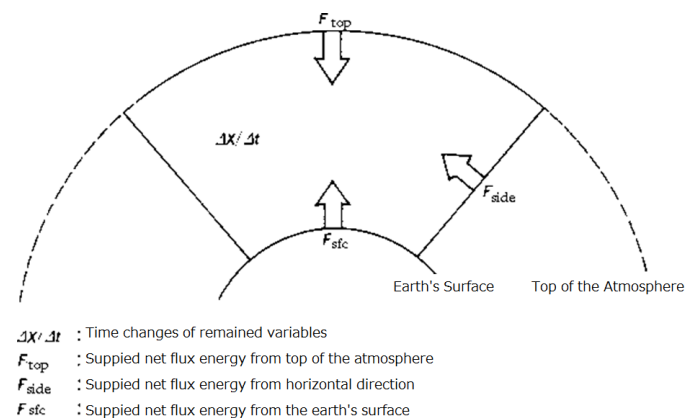


Figure 1 Illustrative view of the assimilation model

The space is divided with meshes in vertical and horizontal directions as shown in Figure 2. In the meshes, variables are maintained and balanced. For instance, input energy to the mesh is totally equal to output energy from the mesh. Thus partial differential equation can be formulated. Because partial differentiation in space is partial differentiation in time, then prediction can be done. Therefore, assimilation can be done. This is called assimilation model.

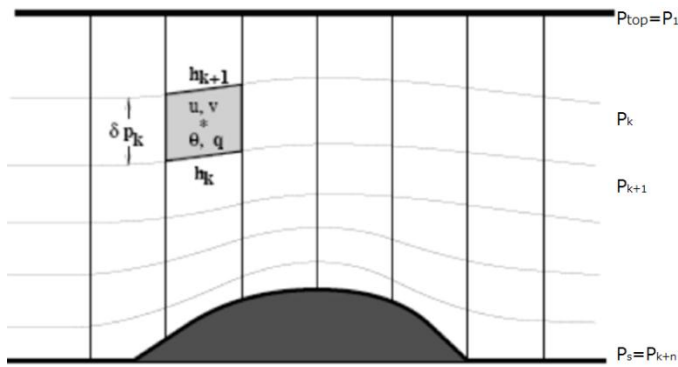


Figure 2 Assimilation model with meshed space and time function of variables.

Variables, input data for assimilation model are obtained from the earth observation satellite data, in particular, infrared sounder as well as microwave sounder data. Table 1 and 2 shows example of the input data, variables, for assimilation model derived from the specific sensor data.

Table 1 Example of the variables, input data for assimilation model and data sources mainly from the satellite data.

Conventional Data	
Radiosondes	u, v, T, q, P _s
PIBAL winds	u, v
Wind profiles	u, v
Conventional, ASDAR, and MDCRS aircraft reports	u, v, T
NEXRAD radar winds	u, v
Dropsondes	u, v, T, P _s
PAOB	P _s
GMS, Meteosat, cloud drift IR and visible winds	u, v
MODIS clear sky and water vapor winds	u, v
GOES cloud drift IR winds	u, v
GOES water vapor cloud top winds	u, v
Surface land observations	P _s
Surface ship and buoy observations	u, v, T, q, P _s
SSM/I	Rain rate, wind speed
TMI	Rain rate
QuikSCAT	u, v

Table 2 Example of sensor name and satellite name of the data sources for assimilation model.

Satellite Data	
TOVS 1b Radiances	AMSU-A: N15, N16, N18 AMSU-B: N15, N16, N17 MHS: N18 HIRS2: TIROS-N, N6, N7, N8, N9, N10, N11, N12, N14 HIRS3: N16, N17 HIRS4: N18 MSU: TIROS-N, N6, N7, N8, N9, N10, N11, N12, N14 SSU: TIROS-N, N6, N7, N8, N9, N10, N11, N14
EOS/Aqua Level 1b Radiances	AIRS (150 channels), AMSU-A
SSM/I radiances	DMSP-8, DMSP-10, DMSP-11, DMSP-13, DMSP-14, DMSP-15 (7 channels)
GOES sounder T _B	GOES-08, GOES-10, GOES-12 Channels 1-18
SEBUV2 ozone (Version 8 retrievals)	Nimbus 7, NOAA 9, 11, 14, 16, 17

The variables are formed time series of spatially aligned three dimensional geophysical data as shown in Figure 3. Layered two dimensional geophysical data, atmospheric pressure, air temperature, relative humidity, etc. aligned along with time direction. Using assimilation model with satellite data derived geophysical data as input data, such this four dimensional geophysical data are created.

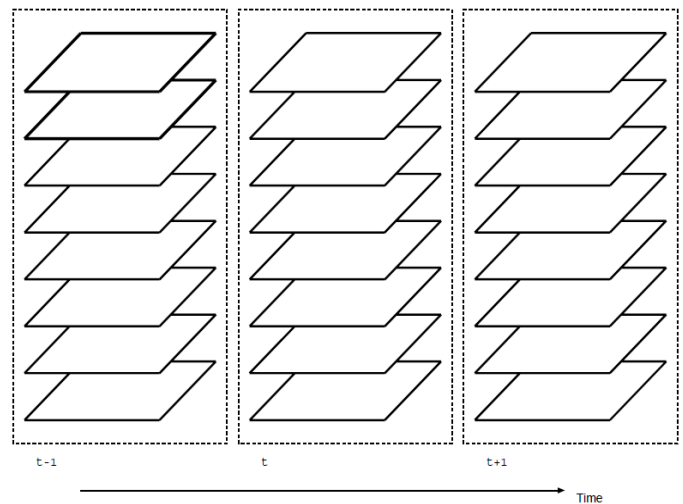


Figure 3 Four dimensional assimilation data

B. Wavelet Analysis Based Problem Solving Method and Change Detections in Space and Time Domains

One dimensional wavelet transformation can be expressed in equation (1).

$$F = Wf \tag{1}$$

where f denotes geophysical data in space or time domain while W denotes wavelet transformation matrix. Since, $[G_{ijk}]^t = G_{kji}$, then three dimensional wavelet transformation can be written as follows,

$$F = [W_k [W_j [W_i G_{ijk}]^t]]^t \tag{2}$$

Not only three dimensional wavelet transformation, but also n dimensional wavelet transformation can be expressed as follows,

$$F = [W_n [W_{n-1} [\dots [W_1 G_{12\dots n}]^t] \dots]]^t \tag{3}$$

This wavelet transformation is called as decompositions.

One of the specific features of the wavelet transformation is that the original geophysical data G can be reconstructed with the calculated wavelet frequency components, F , perfectly. This reconstruction processes is called inverse wavelet transformation.

Also, wavelet analysis is useful for solving integral equations as well as partial differential equation []. Therefore, wavelet analysis based method can be used for solving the partial differential equations in the assimilation model.

Furthermore, wavelet analysis based Multi Resolution Analysis: MRA is used for change detections in space and time domains. Using MRA, wavelet frequency components are calculated. If the reconstruction is applied to the frequency components without low frequency component, then spatial and time changes are extracted because the reconstruction is made with high frequency components only.

C. Proposed MRA Based Change Detection Method

Change detection performance can be evaluated as follows,

$$J1(\alpha) = \sqrt{\frac{\sum_t \sum_z \sum_y \sum_x (d_\alpha^*(x, y, z, t) - d(x, y, z, t))^2}{X \times Y \times Z \times TIME}}$$

$d(x, y, z, t)$: Observation Data
 $d_\alpha^*(x, y, z, t)$: Reconstructed Data (CR= α)
 $X, Y, Z, TIME$: Amount of Data
 $X = 32, Y = 32, Z = 8, TIME = 12$ (4)

where J1 denote cost function which represent change detection performance. Namely, Root Mean Square: RMS difference between the original and the reconstructed images is a good measure for evaluation of change detection performance. The procedure of the proposed method is as follows,

- (1) Wavelet transformation is applied to the original image,
- (2) Wavelet frequency component, coefficients are sorted in accordance with the coefficient values,
- (3) Remove the first n coefficients form the minimum
- (4) Reconstruct image with the rest of coefficients of component
- (5) Determine an optimum n of which the changes of RMS difference is saturated

III. EXPERIMENTS

A. Change Detection Using Subtraction

One of the examples of the change detection in space and time domains is shown. That is cloud movement analysis with Geostationary Meteorological Satellite: GMS imagery data. Figure 4 shows (a) original time series of GMS imagery data, (b) binalized images, and (3) detected changes in space and time domains by using subtraction between adjacent binalized images Also change detection and object movement analysis can be done with optical flow model as well. Thus the clouds movement can be analyzed. It is sensitive to the threshold processes. Namely, detected changes depend on binalized results.

B. Change Detection of Air Temperaure and Relative Humidity Profiles in Space and Time Domains Using the Aforementioned Proposed MRA Based Method

One of examples of change detections based on the proposed wavelet MRA analysis based method is shown. Figure 5 shows monthly average of the relative humidity profiles which were acquired in February and August in 1992. The profile can be formed with 8 layers in altitude direction, z, and with 1 degree resolution in horizontal directions, x and y. Also the relative humidity ranges from 0.005 to 20 g/Kg.

Change detection in time domain can be done with the reconstruction with all wavelet frequency components except LLLL component. LLLL denotes low frequency components of x, y, z, and time directions. However, it is quite obvious that the resultant image shows detected changes in space and time domains as shown in Figure 5 (c).

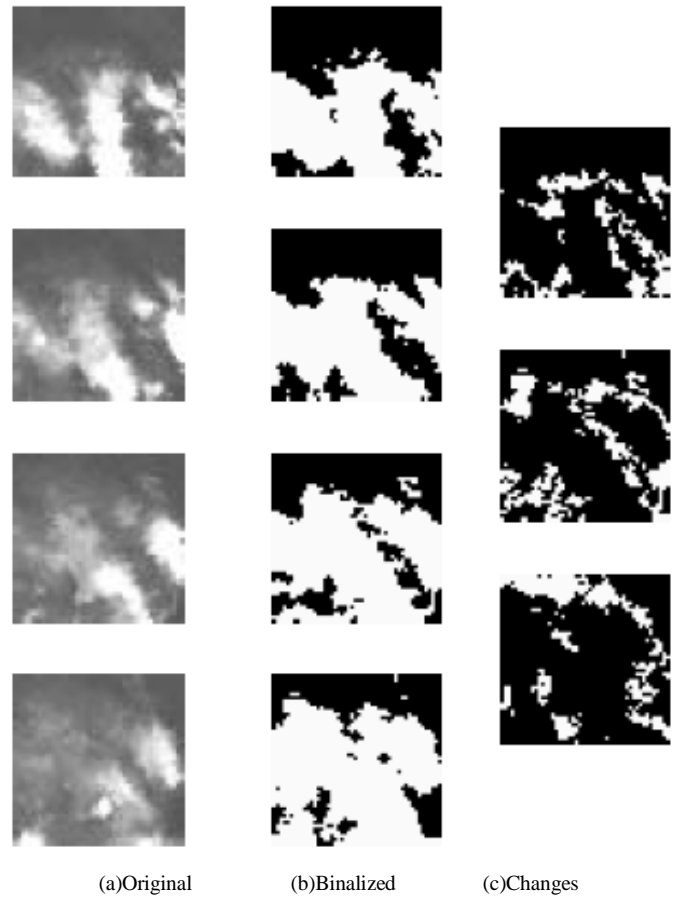
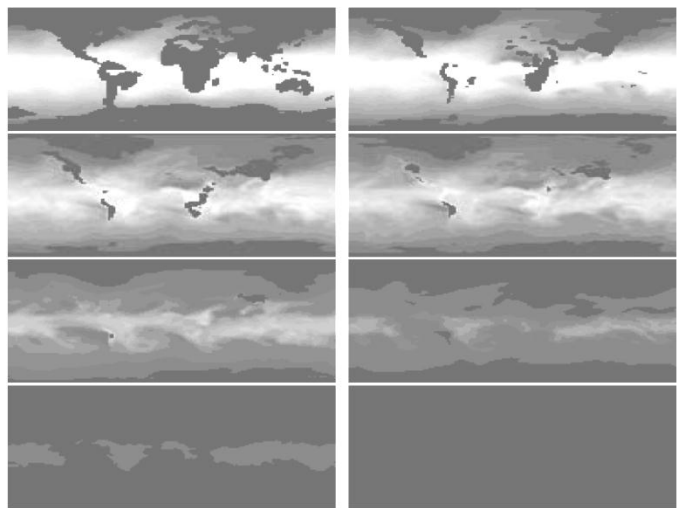
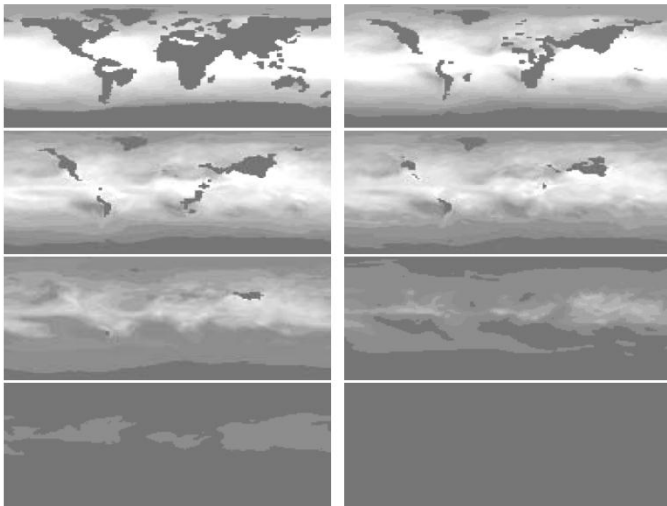


Figure 4 Example of change detection in space and time domains

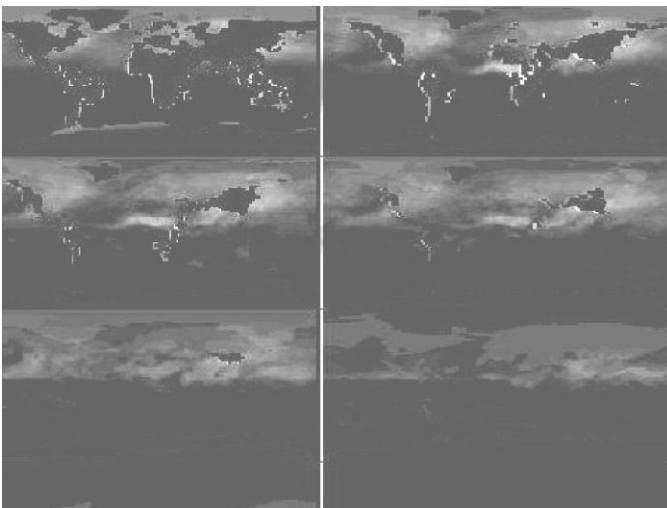
Also, the resultant image seems that not so good performance in terms of detected changes in the image. Then reconstruction with all wavelet frequency components except LLLL component together with the first n wavelet coefficients is tried. In this connection, 100% - α % of wavelet coefficients are removed. α corresponds to data compression ratio.



(a)Relative humidity in August 1992



(b)Relative humidity in February 1992



(c)Detected changes with all frequency components except LLLL component

Figure 5 Detected changes by means of the MRA based method, reconstruction of original image without LLLL component

C. Evaluation of Data Compression Ratio

As the aforementioned in the previous section, change detection performance is highly correlated with the RMS difference between the original and the reconstructed images. Therefore, change detection performance for the proposed method is evaluated with calculation of RMS difference as a function of data compression ratio.

Almost homogeneous ocean area is selected for the evaluation. In this Pacific Ocean area which is shown in Figure 6, relative humidity is relatively homogeneous. Relation between RMS difference and data compression ratio is shown in Figure 7. Although RMS error for the data compression ratio of 100% (which means no data compression is applied) is obviously zero, it is getting large in accordance with decreasing of data compression ratio. For instance, RMS error for data compression ratio of 10% is 0.4. This implies that 0.4 of RMS error has to be accepted for 1/10 of data compression.



Figure 6 Intensive study area for RMS difference evaluations

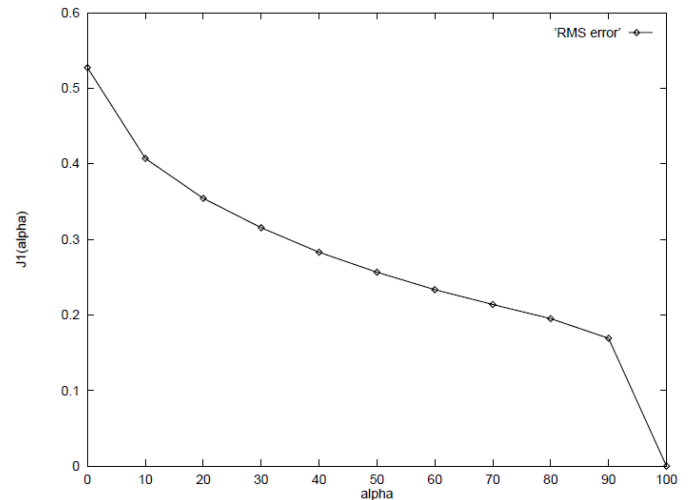


Figure 7 Relation between RMS difference and data compression ratio

IV. CONCLUSION

Method for time change detection of four dimensional assimilation data by means of wavelet analysis is proposed together with spatial change detection with satellite imagery data. The method is validated with assimilation data and the same scale of satellite imagery data. Experimental results show that the proposed method does work well in visually. Also the relation between RMS error and data compression ratio is clarified.

ACKNOWLEDGMENT

The author would like to thank Dr. Kaname Seto for his effort to conduct experiments.

REFERENCES

- [1] R. Daley, Atmospheric data analysis, Cambridge University Press, 1991.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [2] Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc (1997) Unified Notation for Data Assimilation: Operational, Sequential and Variational Journal of the Meteorological Society of Japan, vol. 75, No. 1B, pp. 181–189
- [3] John M. LEWIS; S. Lakshmivarahan, Sudarshan Dhall, "Dynamic Data Assimilation : A Least Squares Approach", Encyclopedia of Mathematics and its Applications 104, Cambridge University Press, 2006

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science, and Technology of the University of Tokyo from 1974 to 1978 also was with National Space Development Agency of Japan (current JAXA) from 1979 to 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He was appointed professor at Department of Information Science,

Saga University in 1990. He was appointed councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was also appointed councilor of Saga University from 2002 and 2003 followed by an executive councilor of the Remote Sensing Society of Japan for 2003 to 2005. He is an adjunct professor of University of Arizona, USA since 1998. He also was appointed vice chairman of the Commission "A" of ICSU/COSPAR in 2008. He wrote 30 books and published 332 journal papers.

Method for Image Source Separation by Means of Independent Component Analysis: ICA, Maximum Entropy Method: MEM, and Wavelet Based Method: WBM

Kohei Arai¹

Graduate School of Science and Engineering
Saga University
Saga City, Japan

Abstract—Method for image source separation based on Independent Component Analysis: ICA, Maximum Entropy Method: MEM, and Wavelet Based Method: WBM is proposed. Experimental results show that image separation can be done from the combined different images by using the proposed method with an acceptable residual error.

Keywords—Blind separation; image separation; cucktail party effect; ICA; MEM; wavelet analysis.

I. INTRODUCTION

There are some technologies and systems which allow separate the specific speaker from the mixed image data which is acquired at some noisy circumstances. The related application studies are conducted, in particular, for TV meeting environment, image recognition systems, and digital hearing aid system etc. In particular, the microphone array system as well as Independent Component Analysis (ICA) base approach [1] is focused. Microphone array allows enhance a target image from the mixed images suppressing noises and taking into account the phase difference among the imagesources which corresponds to the distance between the microphone and the location of the imagesources. There is a delay sum [2] and an adaptation [3] types array microphone systems. These types of array microphone allow direct the beam to the desired direction of the target of interest.

ICA is the method which allows the blind separation based on the imagesources are isolated each other. ICA based method configures reconstruction filter maximizing Kullback-Leibler Divergence [4] for separation of target imagesources in concern from the acquired images [5], [6].

The blind separation method with entropy maximization rule is proposed [7]. In order to improve separability among the possible imagesources, high frequency component derived from the wavelet based Multi Resolution Analysis (MRA) [8],[9] is used in the entropy maximization rule utilized method. Due to the fact that the MRA based separability improvement is not good enough, further improvement is required.

Blind separation method which is proposed here is based on the MRA based separability improvement. A single level of MRA is not good enough for characteristics enhancement of each imagesource. Therefore, the level is considered as a parameter for MRA utilized blind separation [10]. An appropriate level is found for improvement of separability then blind separation is applied. It is found that the proposed method can achieve 4 to 8.8% of separability improvement for the case of the number of speakers is 2, 4 and 8 [11].

The following section describes the proposed method followed by some experiments with the mixed images. Then conclusions and some discussions is followed.

II. PROPOSED MTHEOD

A. Image Mikxing Model

Assuming the original image, $x_i(t)$ is mixed with several images, for instance $s_1(t)$ and $s_2(t)$ as shown in equation (1).

$$x_i(t) = a_{i1}s_1(t) + a_{i2}s_2(t) \quad (1)$$

where a_{i1} and a_{i2} are weighting factor, mixing ratio. Also it is assumed that $s_1(t)$ and $s_2(t)$ are mutually independent. Although this is an example of image mixing model for just two images, it is possible to expand the number of images which are to be mixed together.

B. Maximum Entropy Method

Maximum Entropy Method: MEM is used for learning processes for maximizing combined entropy in two layered neural network

$$y_i = g(\sum_{k=1}^2 w_{ik} x_k - \theta_i) \quad (2)$$

$$g(v) = 1 - e^{-v} / (1 + e^{-v}) \quad (3)$$

where x, y notes input and output signals, or images while w denotes weighting coefficients of two layered neural network. θ denotes a threshold. $g(v)$ is called sigmoid function.

The combined entropy can be expressed with the equation (4).

$$H(y) = \langle \ln(|J|) \rangle - \langle \ln(p(x)) \rangle \quad (4)$$

where J denotes Jacobian matrix as shown in equation (5)

$$J = \det \begin{pmatrix} \partial y_1 / \partial x_1 & \partial y_1 / \partial x_2 \\ \partial y_2 / \partial x_1 & \partial y_2 / \partial x_2 \end{pmatrix} \quad (5)$$

Because the second term of the equation (4) is constant, the following equation (6) has to be maximized.

$$I = - \langle \ln(|J|) \rangle \quad (6)$$

C. Steepest Descending Method

In order to maximize the equation (6), Steepest Descending Method: SDM is used. Updating equations for w and θ can be expressed with equations (7) and (8).

$$w_{ik}^{new} = w_{ik}^{old} - \gamma \partial I / \partial w_{ik} \quad (7)$$

$$\theta_i^{new} = \theta_i^{old} - \gamma \partial I / \partial \theta_i \quad (8)$$

More precisely, updating equations can be re-written by equation (9) and (10).

$$W^{new} = W^{old} + \gamma \langle [W^T]^{-1} + 2yx^T \rangle \quad (9)$$

$$\theta^{new} = \theta^{old} + \gamma \langle 2y \rangle \quad (10)$$

Thus the original images are separated from the combined image with the proposed method.

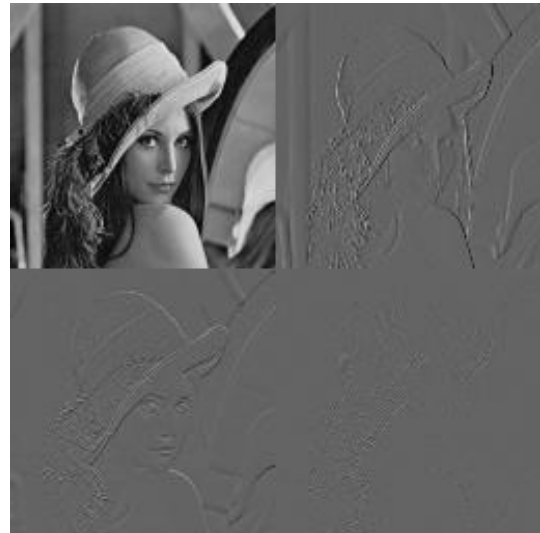
D. Discrete Wavelet Transformation

In order to separate several original images from the given combined images, Discrete Wavelet Transformation: DWT is used. DWT allows decompose images with high and low wavelet frequency components orthogonally. There are many orthogonal base functions for DWT. One of the based functions is Haar function. Figure 1 shows an example of the original image of Lena in the SIDBA standard image database, and decomposed image through WT. In the decomposed image, LL image, low frequency component in horizontal direction and low frequency component in vertical direction is situated at the top left corner while LH, HL, and HH components are situated at the top right, the bottom left, and the bottom right corners, respectively.

Histogram of the low frequency component is shown in Figure 2 (a) while that of the high frequency component is shown in Figure 2 (b). This histogram is for the decomposed image of Lena image. The histogram (a) is for LL component while the histogram (b) is for HH component. The histogram of the high frequency component looks similar to normal distribution and is mainly concentrated at the wavelet frequency component ranges from 0 to 10 wavelet frequency. On the other hand, Histogram of the LL component does not look like normal distribution.



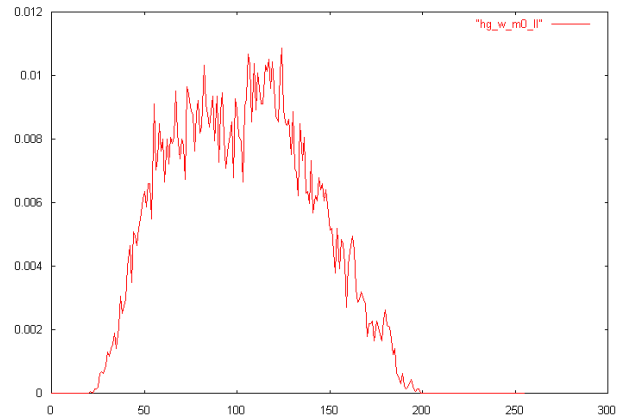
(a)Original image



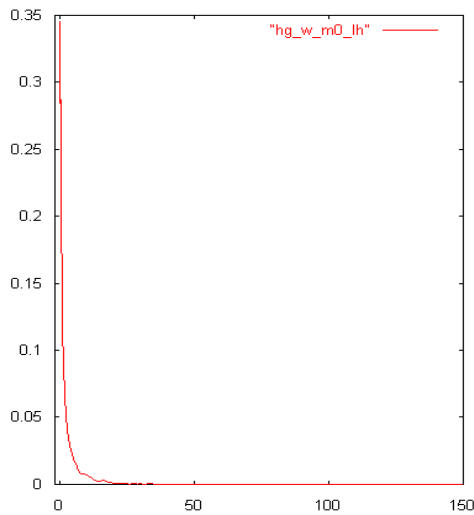
(b)Decomposed image

Figure 1 Example of original and decomposed images with DWT.

These histogram characteristics are common to original image and differ from each other depending on natures of the original images. Therefore, it is possible to separate original images using the difference of histograms of the decomposed image derived through DWT.



(a)Low Frequency Component



(b)High Frequency Component

Figure 2 Histograms of high and low frequency components of the decomposed image after DWT

III. EXPERIMENTS

A. Original Images Used

Figure 3 shows two original images, Lena, and Barbara in the same standard image database, SIDBA.



(a)Lena



(b)Barbara

Figure 3 Two original images used for experiments

Also examples of mixed images are shown in Figure 4. Mixing ratios of Figure 4 (a) and (b) are different. Mixing ratio of Figure 4 (a) is 90% of Lena image and 10% of Barbara image while that of Figure 4 (b) is 10% of Lena image and 90% of Barbara image. Image As shown in Figure 4, image defects are found on the mixed, combined images.

Figure 5 shows DWT applied images of Figure 4 of combined images.



(a)Combined image with 90% of Lena and 10% of Barbara



(b) Combined image with 90% of Barbara and 10% of Lena
Figure 4 Combined images with the different mixing ratios



(b) DWT image of Figure 3 (b) of combined image

Figure 5 DWT images of Figure 4.



(a) DWT image of Figure 3 (a) of combined image

B. Separated Images with the Different Mixing Ratios

Mixed image between Lena and Barbara images with the different mixing ratios, 50%, 30%, and 10% are created. Using the proposed method, separation of each original image is attempted. Resultant images are shown in Figure 6, 7, and 8 for the mixing ratios, 50%, 30%, and 10%, respectively. As shown in these Figures, image separation performance depends on the mixing ratio. It is difficult to separate image when the mixing ratio is around 50%. Meanwhile, image separation performance is getting better in accordance with decreasing of the mixing ratio. In particular, the separated images for the mixing ratio of 10% are almost perfect, look extremely similar to the original images.



(a)



(b)

Figure 6 Separated images in the case of which mixing ratio is 50%.



(b)

Figure 7 Separated images in the case of which mixing ratio is 30%.



(a)



(a)



(b)

Figure 8 Separated images in the case of which mixing ratio is 10%.

IV. CONCLUSION

Method for image source separation based on Independent Component Analysis: ICA, Maximum Entropy Method: MEM, and Wavelet Based Method: WBM is proposed. Experimental results show that image separation can be done from the combined different images by using the proposed method with an acceptable residual error.

It is found that the image separation performance depends on the mixing ratio. It is difficult to separate image when the mixing ratio is around 50%. Meanwhile, image separation performance is getting better in accordance with decreasing of the mixing ratio. In particular, the separated images for the mixing ratio of 10% are almost perfect, look extremely similar to the original images.

ACKNOWLEDGMENT

The author would like to thank Takeshi Yoshida for his effort to the experiments.

REFERENCES

- [1] A. Hyvarinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley and Sons, 2001.
- [2] H. Nomura, Y. Kaneda, N. Kojima, Near Field Types of Microphone Array, Journal of the Acoustical Society of Japan, 53, 2, 110-116, 1997.
- [3] Y. Kaneda, Adaptive Microphone Array, Journal of the Institute of Electronics, Information and Communication Engineers, J71-B-II, 11, 742-748, 1992.
- [4] M. Takagi and H. Shimoda Edt. K.Arai et al., Image Analysis Handbook, Tokyo-Daigaku-Shuppankai Pub. Co. Ltd., 1991.
- [5] C.Jutten,J.Herault, Blind separation of sources, Part I: An adaptive algorithm based on neuron, Signal Processing, 24, 1-10, 1991.
- [6] S. Amari, N. Murata, Independent Component Analysis -A New Method for Multi-Variate Data Analysis, Science Pub. Co. Ltd., 2002.
- [7] S. Morishita, S. Miyano Edt., K. Nijima, Extraction of hidden image signals by means of blind separation and wavelets, Kyoritsu Shuppan Pub. Co. Ltd., 201-206, 2000.
- [8] K.Arai, Fundamental Theory on Wavelet Analysis, Morikita Shuppan Pub. Co. Ltd., 2000.
- [9] K.Arai, Self Learning on Wavelet Analysis, Kindai-Kagakusha Shuppan Co. Ltd., 2003.
- [10] K. Arai, T. Yoshida, Speaker separation based on blind separation method with wavelet transformations, Journal of the Visualization Society of Japan, 26, Suppl.1, 171-174, 2006.
- [11] K.Arai, Blind separation, IJRCS

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada.

He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission "A" of ICSU/COSPAR since 2008. He wrote 30 books and published 322 journal papers.

Multifinger Feature Level Fusion Based Fingerprint Identification

Praveen N

Research Fellow Audio and Image Research
Laboratory Department of Electronics Cochin
University of Science and Technology
Cochin, Kerala, India.

Tessamma Thomas

Professor
Department of Electronics
Cochin University of Science and Technology
Cochin, Kerala, India.

Abstract— Fingerprint based authentication systems are one of the cost-effective biometric authentication techniques employed for personal identification. As the data base population increases, fast identification/recognition algorithms are required with high accuracy. Accuracy can be increased using multimodal evidences collected by multiple biometric traits. In this work, consecutive fingerprint images are taken, global singularities are located using directional field strength and their local orientation vector is formulated with respect to the base line of the finger. Feature level fusion is carried out and a 32 element feature template is obtained. A matching score is formulated for the identification and 100% accuracy was obtained for a database of 300 persons. The polygonal feature vector helps to reduce the size of the feature database from the present 70-100 minutiae features to just 32 features and also a lower matching threshold can be fixed compared to single finger based identification.

Keywords- fingerprint; multimodal biometrics; gradient; orientation field; singularity; matching score.

I. INTRODUCTION

Personal authentication based on biometric traits is the most common in the current security access technologies. As the criminal/fraudulent activities are increasing enormously, designing high security identification has always been the main goal in the security business. Biometrics deals with identification of people by their physical and/or behavioral characteristics and, so, inherently requires that the person to be identified is physically present at the point of identification. Fingerprints offer an infallible means of personal identification [1]. The large numbers of fingerprint images, which are collected for criminal identification or in business for security purpose, continuously increase the importance of automatic fingerprint identification systems. Most of the automatic fingerprint identification systems can reach around 97% accuracy with a small database and the accuracy of identification is drops down as the size of database is growing up [2, 3]. Also, the processing speed of automatic identification systems decreases if it involves a large number of detection features. Hence feature code template size has to be minimized so that identification may be much easier. A fingerprint is characterized by singularities- which are small regions where ridge lines forms the distinctive shapes: loop, delta or whorl (Fig.1). Singularities play a key role in classification of fingerprints [1, 5] which sets fingerprints into a specific set.



Fig.1. Fingerprint Singularities

Classification eases the searching and in many of the AFIS, classification is the primary procedure adopted [6].

According to Galton-Henry classification scheme [1, 4] there are four common classes of fingerprints: Arch and Tented Arch, Left loop, Right Loop and Whorl (Fig. 2). Cappelli and Maltoni, 2009 [7] studied the spatial distribution of fingerprint singularities and proposed a statistical model for the four most common classes: Arch, Left loop, Right loop and Whorl. The model they proposed gives a clear indication of the fingerprint identity and is used here for identification with a sharp reference position.

Biometric systems that use a single modality are usually affected by problems like noisy sensor data, non-universality and/or lack of distinctiveness of the biometric trait, unacceptable error rates, and spoof attacks [8]. Multibiometric system deals with two or more evidences that are taken from different sources like multiple fingers of the same person,



Fig.2. Fingerprint Types

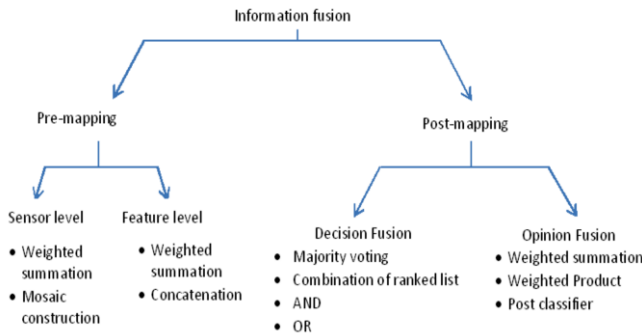


Fig.3. Information fusion

multiple samples of the same instances, multiple sensors for the same biometric, multiple algorithms for representation and matching of multiple traits [9]. Information fusion refers to the consolidating of information or evidences presented by multiple biometric sources [10, 11, 12].

II. FUSION IN BIOMETRICS

Hall and Llinas[13], Ross and Jain[8] have divided information fusion into several categories: sensor level fusion, feature level fusion, score level fusion and decision level fusion. Based on this Sanderson and Paliwal[14] have classified information fusion into pre-mapping fusion, midst-mapping fusion and post-mapping fusion [Fig. 3]. Pre-mapping fusion refers to combining information before any use of classifiers or experts. In midst-mapping fusion, information is combined during mapping from sensor data/feature space into opinion/decision space and in post-mapping fusion, information is combined after the decisions of the classifiers have been obtained. Match score fusion based multibiometric algorithm have been developed by Ross and Jain, 2003[8], Frischholz and Dieckmann, 2000[15], Hong and Jain, 1998[16], Biguin et al., 1997[17], Wang et al., 2003[18], Kumar and Zhang, 2003[19]. Fusion at the match score, rank and decision level have been developed and studied extensively. Feature level fusion, however, is a relatively understudied problem [20].

In the present work, feature level fusion of feature vectors is done by concatenating individual feature vectors of consecutive fingers. Fingerprint baseline, which is defined as the line between the Distal and Intermediate Phalangeal joint line is taken as the reference line [Fig. 4]. This line is detected using correlation technique, singularities are detected using directional field strength and a polygon is formed with

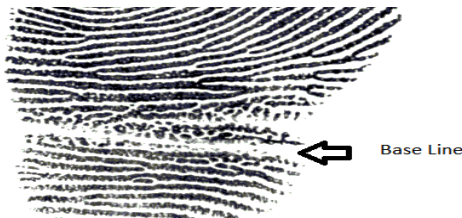


Fig.4. Fingerprint Baseline

singularities and the baseline. For each finger, feature vector is computed as the distance, angle parameters and ridge counts which are concatenated to form the multifinger feature vector. Matching score is formulated to identify the fingerprint. FAR and FRR curves are plotted.

III. DEFINITION OF THE NOVEL FINGERPRINT STRUCTURE AND FEATURE VECTOR FORMATION

In this work, fingerprint singularities are identified and a polygon is formed with the baseline [21] [Fig.5]. The polygon thus formed is invariant to rotation. Feature vector describing the polygon is defined as $F = (d, \theta, A, T, r)^T$ where d is the distance metric, θ is the angle metric, A is the area of the polygon, T is the type of the fingerprint/polygon and r is the ridge counts. The feature vector thus formed is a 16 element vector as:

$$[d_{cc}, d_{cb}, d_{cdr}, d_{abr}, d_{bb}, d_{abl}, d_{cdl}, \theta_c, \theta_{dr}, \theta_{dl}, \theta_{cc}, A, T, r_{cd}, r_{cb}, r_{ab}]^T$$

where $d_{cc}, d_{cb}, \dots, d_{cdl}$ are the distance measures, $\theta_c, \theta_{dr}, \theta_{cc}$ are the convex angle metrics, A is the Area of the polygon formed and r_{cb}, r_{cb} & r_{db} are the ridge counts between core-delta, core-base and delta-base respectively. These are shown in fig.5. Following steps are carried out for constructing the fingerprint polygon:

A. Directional Field Estimation and Strength

Directional field shows the coarse structure or basic shape of a fingerprint [22] which gives the global information about a fingerprint image. It is defined as the local orientation of the ridge-valley structures.

By computing directional field, singularities can be efficiently located. Several methods have been adopted to estimate directional field. [23], [24], [25]. M. Kass and Witkin, [26] introduced the gradient based method and was adopted by fingerprint researchers [27, 28, 29, 30]. This method is

The gradient $\begin{bmatrix} G_x(x, y) \\ G_y(x, y) \end{bmatrix} = \nabla I(x, y) = \begin{bmatrix} \frac{\partial I(x, y)}{\partial x} \\ \frac{\partial I(x, y)}{\partial y} \end{bmatrix}$ vector $[G_x(x, y) \ G_y(x, y)]^T$ is defined as:

$$(1)$$

Where $I(x, y)$ represents the gray-scale image. The directional field is perpendicular to the gradients. Gradients are orientations at pixel-scale whereas directional field describes orientation of ridge-valley structure.

An averaging operation is done on the gradients to obtain the directional field. Gradients cannot be averaged in the local neighborhood as opposite gradients will cancel each other. To solve this problem Kass and Witkin doubled the angle of the gradient vectors before averaging. Doubling makes opposite vectors points in the same

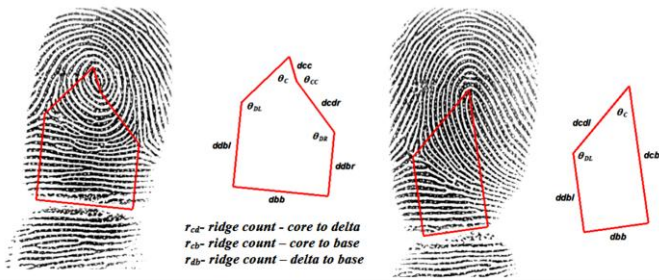


Fig.5. Fingerprint Polygon

direction and will reinforce each other while perpendicular gradients will cancel each other. After averaging, the gradient vectors have to be converted back to their single-angle representation.

The gradient vectors are estimated first in Cartesian coordinate system and is given by $[G_x, G_y]$. For the purpose of doubling the angle and squaring the length, the gradient vector is converted to polar system, which is given by

$$[\rho\phi]^T \text{ where } -\frac{1}{2}\pi < \phi \leq \frac{1}{2}\pi$$

$$\begin{bmatrix} \rho \\ \phi \end{bmatrix} = \begin{bmatrix} \sqrt{G_x^2 + G_y^2} \\ \tan^{-1} G_y / G_x \end{bmatrix} \quad (2)$$

The gradient vector is converted back to its Cartesian as:

$$\begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \rho \cos \phi \\ \rho \sin \phi \end{bmatrix} \quad (3)$$

The average squared gradient $\begin{bmatrix} \overline{G_{s,x}} \\ \overline{G_{s,y}} \end{bmatrix}$ is given by

$$\begin{bmatrix} \overline{G_{s,x}} \\ \overline{G_{s,y}} \end{bmatrix} = \begin{bmatrix} \sum_W G_x^2 - G_y^2 \\ \sum_W 2G_x G_y \end{bmatrix} = \begin{bmatrix} G_{xx} - G_{yy} \\ 2G_{xy} \end{bmatrix} \quad (4)$$

where

$$\begin{aligned} G_{xx} &= \sum_W G_x^2 \\ G_{yy} &= \sum_W G_y^2 \\ G_{xy} &= \sum_W G_x G_y \end{aligned} \quad (5)$$

are estimates for the variances and cross covariance of G_x and G_y , averaged over the window W . The average gradient direction ϕ is given by:

$$\phi = \frac{1}{2} \angle(G_{xx} - G_{yy}, 2G_{xy}) \quad (6)$$

where $\angle(x,y)$ is defined as:



Fig.6. Fingerprint and Directional Field

$$\angle(x, y) = \begin{cases} \tan^{-1}(y/x) & x \geq 0 \\ \tan^{-1}(y/x) + \pi & \text{for } x < 0 \wedge y \geq 0 \\ \tan^{-1}(y/x) - \pi & x < 0 \wedge y < 0 \end{cases}$$

Directional field image obtained is shown in Fig. 6.

B. Singularity Detection and Fingerprint Classification

Singularities are the vertices of the fingerprint polygon. The most common method used for singularity detection is by means of Poincaré index proposed by Kawogoe and Tojo, 1984[31].

Poincaré index is given by

$$PP_{GC}(i, j) = \sum_{k=0 \dots 7} \text{angle}(d_k, d_{(k+1) \bmod 8}) \quad (7)$$

Where G is the field associated with the fingerprint orientation image, D is the closed path defined as an ordered sequence; d is the directional field of individual blocks around region of interest (Table 1).

TABLE 1. POINCARÉ INDEX COMPUTATION SCHEME

d_2	d_3	d_4
d_1	$[i, j]$	d_5
d_0	d_7	d_6

Poincaré index method cannot accurately detect the singular points for noisy or low quality fingerprints and for singular points in arch fingerprints and some of the tented arch fingerprints [32]. *Coherence*, which gives the strength of the orientation, measures how well all squared gradient vectors share the same orientation [26]. In this work, singularities are

2. Compute matching score,

$$M = \begin{cases} \frac{Th - \sum_{n=1}^N |I_n - F_n| * w_n}{Th} & \text{if } \sum_{n=1}^N |F_n - I_n| * w_n < Th \\ 0 & \text{otherwise} \end{cases}$$

Th is the Threshold value assigned, w_n is the weight, I_n and F_n are the Input image and Template metrics, N is the feature vector length.

3. If $M \leq Th$, fingerprint matches, where $0 \leq Th \leq 1$.

VI. IMPLEMENTATION OF THE ALGORITHM

A. Database used

In our work consecutive fingerprint images (forefinger and middle finger) have been acquired using a fingerprint scanner with 500 dpi resolution and image size of 600 X 600 pixels. Fingerprint samples of about 300 persons were collected and features were extracted and stored.

B. Implementation

Individual fingerprints are segmented to detect the baseline and to generate the feature vectors. Baseline is captured clearly for individual fingerprints and the polygon is drawn for whorl, left loop, right loop and arch classes of the input images and the features from the polygon are evaluated. Classification of fingerprints is done for each fingerprint as per the classification scheme and the templates are formed and stored [Table 2].

C. Results and Discussion

About 300 fingerprint pairs were taken and the features were extracted and stored as template data. Another sample set of 300 fingerprint pairs of same persons were taken as test set. Match score has been calculated for each fingerprint in the test set with the template. Box plot, which is a statistical plot of the score distribution, is shown in fig.10. A genuine fingerprint is one which is supposed to match with the same fingerprint template in the template data. The genuine distribution shows a median of about 0.92 and the whiskers ranging between 0.72 to 1. An imposter in one whose fingerprint does not matches with the template data. The imposter distribution shows a median of 0.56 with the whiskers ranging between 0 and 0.71. Hence fixing a matching score threshold between 0.72 and 0.71 can identify all the fingerprints with 100% accuracy. The Receiver Operator Characteristics (ROC) graph is shown in fig.11. The False Acceptance Rate (FAR) is a measure of how many imposter users are falsely accepted into the system as "genuine" users is plotted by varying the threshold from 0 to 1. The False Rejection Rate (FRR) is a measure of how many genuine users are falsely rejected by the system as "imposters" is also calculated for various thresholds and is plotted. The Equal Error Rate (EER) is defined as the condition at which FAR=FRR as in Figure 11, is approximately equal to zero at a threshold of $Th = 0.715$. For this threshold, fingerprints are identified with 100% accuracy. Fig.12 shows the ROC correspond to single finger based identification which shows a high threshold can only identify the fingerprint with 100 % accuracy.

TABLE 2. FINGERPRINT TEMPLATE

d_{cc}	d_{cb}	d_{cdr}	d_{dbr}	d_{bb}	d_{dbl}	d_{cdl}	θ_C	θ_{DR}	θ_{DL}	θ_{CC}	A	T	r_{cd}	r_{cb}	r_{db}
0.00	234.47	0.00	0.00	110.74	181.77	122.70	60.13	0.00	119.88	0.00	23049.00	2	10.67	18.67	15.33
0.00	219.56	0.00	0.00	86.41	151.14	110.49	44.77	0.00	135.18	0.00	16055.00	2	11.00	21.00	13.33
0.00	258.32	0.00	0.00	141.61	144.20	179.13	50.25	0.00	129.55	0.00	27670.75	2	17.50	16.75	10.50
0.00	256.01	181.95	120.13	121.06	0.00	0.00	41.82	138.10	0.00	0.00	22805.50	1	13.67	15.17	8.33
0.00	252.18	0.00	0.00	109.93	142.25	155.49	44.95	0.00	135.15	0.00	21605.67	2	13.33	21.50	14.50
0.00	237.91	0.00	0.00	12.02	187.32	52.30	13.33	0.00	166.66	0.00	2567.25	2	3.00	15.50	14.75
0.00	253.24	0.00	0.00	44.01	218.95	56.31	52.07	0.00	127.82	0.00	10451.00	2	7.00	20.00	16.50
0.00	317.00	0.00	0.00	152.00	124.00	245.67	17.12	0.00	162.88	0.00	33516.00	2	17.00	18.00	6.00
0.00	250.12	0.00	0.00	101.24	174.45	126.45	47.78	0.00	132.20	0.00	21503.25	2	11.00	20.75	14.00
0.00	231.18	0.00	0.00	43.44	172.38	73.14	36.61	0.00	143.34	0.00	8782.83	2	5.00	21.50	16.00
0.00	281.00	0.00	0.00	125.00	105.00	193.72	21.79	0.00	158.21	0.00	25875.00	2	16.00	21.50	9.00
0.00	271.20	0.00	0.00	124.81	103.71	197.70	29.88	0.00	150.28	0.00	24235.75	2	16.00	17.50	9.25
0.00	297.74	182.66	129.60	71.56	0.00	0.00	23.05	156.96	0.00	0.00	15289.25	1	15.00	20.50	8.00
0.00	230.95	158.13	121.03	113.58	0.00	0.00	45.92	134.14	0.00	0.00	19952.33	1	13.67	17.00	9.33
0.00	249.02	102.70	171.14	66.85	0.00	0.00	37.23	142.76	0.00	0.00	14043.33	1	8.33	18.17	12.17
0.00	233.79	91.84	184.95	77.24	0.00	0.00	55.65	124.32	0.00	0.00	16181.00	1	10.33	23.33	14.00
163.77	129.25	221.05	114.21	101.18	0.00	0.00	168.02	142.36	0.00	154.31	20160.00	5	12.00	14.00	16.00

VII. CONCLUSION

Multifinger fusion based fingerprint identification is presented here. Singularities of consecutive individual fingerprints were found out using coherence computed via directional field strength. Fingerprint polygon was constructed for each fingerprints with the baseline detected and feature vectors were concatenated to form a 32 element vector. A distance based matching score was formulated and was tested with an accuracy of 100% detection for a database of 300 candidates.

REFERENCES

- [1] David Maltoni, Dario Maio, Anil K Jain, Salil Prabhakar, Handbook of fingerprint recognition, 2nd Ed. Springer, 2005.
- [2] A.K. Jain, L. Hong and R. Bolle, "On-line fingerprint verification", IEEE Trans. Pattern Anal. Mach. Intell. 19 4 (1997), pp. 302-314.
- [3] A.K. Jain, S. Prabhakar and S. Pankanti, "Filterbank-based fingerprint matching", IEEE Trans. Image Process. 9 5 (2000), pp. 846-859.
- [4] Nalini Ratha, Ruud Bolle Editors, Automatic Fingerprint Recognition Systems, Springer, 2003.
- [5] R. Cappelli and D. Maio, "The state of the Art in Fingerprint Classification", Automatic Fingerprint Recognition Systems, N. Ratha and R. Bolle, eds., Springer, 2004.
- [6] R. Cappelli, D. Maio and D. Maltoni, "Indexing Fingerprint Databases for Efficient 1: N Matching", Proc. Sixth International Conf. Control, Automation, Robotics and Vision, Dec. 2000.
- [7] Raffaele Cappelli and Davide Maltoni, "On the Spatial Distribution of Fingerprint Singularities", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 31, No. 4, April 2009 pp- 742-748.
- [8] A Ross, A Jain, "Information fusion in biometrics", Pattern Recogn. Lett. 24 (2003)2115-2125
- [9] Jain A K, Nandakumar K, and Ross A, "Score Normalization in Multimodal Biometric Systems", Pattern Recognition, 38(12), 2270-2285
- [10] Arun A Ross, Karthik Nandakumar, Anil K Jain, Handbook of Multibiometrics, Springer, 2006.
- [11] Biometrics in the Age of Heightened Security and Privacy. Available at http://www.itl.nist.gov/div895/isis/bc/bc2001/EDIT_FINAL_DR.ATTICK.pdf
- [12] Most M, "Battle of Biometrics", Digital ID World Magazine, 2003, Pages 16-18.
- [13] D L Hall, J Llinas, "Multisensor data fusion", D L Hall, J Llinas (Eds.), Handbook of Multisensor Data Fusion, CRC Press, 2001, pp 1-10.
- [14] C. Sanderson, K.K. Paliwal, "Identity Verification Using Speech and Face Information", Digital Signal Processing, Vol. 14, No. 5, 2004, pp. 449-480.
- [15] R Frischholz and U Dieckmann, Biod: "A multimodal biometric identification system", IEEE Computer, 33(2), pp 64-68, 2000.
- [16] L Hong and A K Jain, "Integrating faces and fingerprints for personal identification", IEEE Trans. on Pattern Analysis and Machine Intelligence 20(12), p 1295-1307, 1998.
- [17] E Bigun, J Bigun, B Due and S Fischer, "Expert conciliation for multi modal person authentication by Bayesian statistics", Proc. of Int'l Conf. on Audio and Video based Person Authentication, pp 311-318, 1997.
- [18] Y.Wang, T Tan and A. K. Jain, "Combining face and iris biometrics for identity verification", Proc. of Int'l Conf. on Audio and Video based Person Authentication, pp 805-813, 2003.
- [19] A Kumar and D Zhang, "Integrating palmprint with face for user authentication", Workshop on Multi Modal User Authentication (MMUA), pp 107-112, 2003.

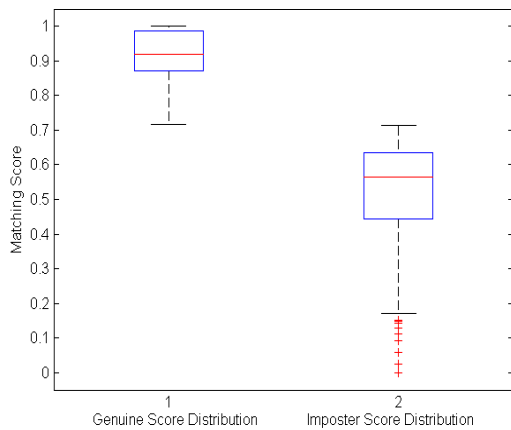


Fig.10. Score Distribution – Multifinger

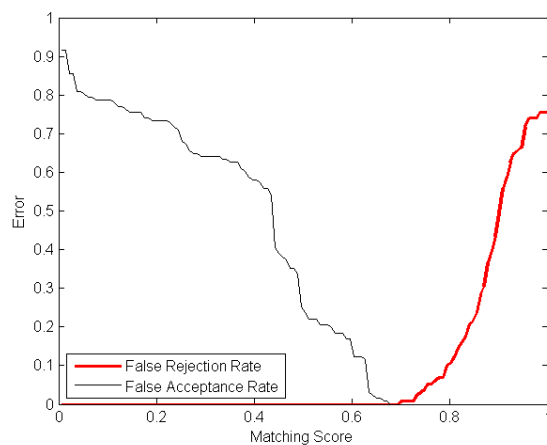


Fig.11. FAR-FRR Graph- Multifinger

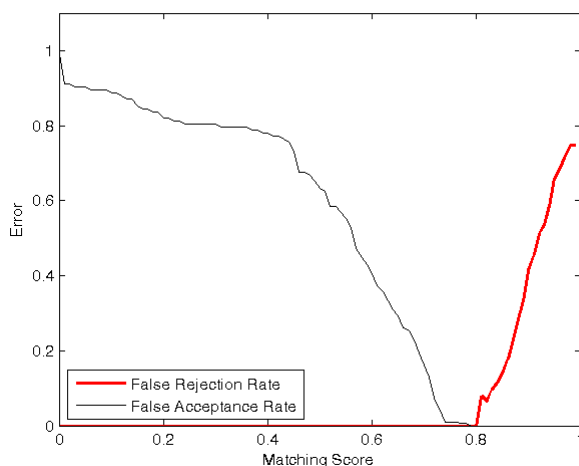


Fig.12 FAR-FRR Graph- Single finger based

- [20] Arun Ross and Rohin Govindarajan, "Feature Level Fusion Using Hand and Face Biometrics", Proc. of SPIE Conference on Biometric Technology for Human Identification II, Vol 5779, pp 196-204, 2005.
- [21] N Praveen and Tessamma Thomas, 'Singularity Based Fingerprint identification', International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 2, No. 5, pp 1055-1059, October 2011.
- [22] Asker M Bazen and Sabih H. Gerez, "Systematic Methods for the Computation of the Directional Fields and Singular Points of Fingerprints", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, July, 2002 pp 905-919.
- [23] G. A. Dretsand G. G. Liljenstr'Om, "Fingerprint Sub classification: A Neural Network Fingerprint Classification", Intelligent Biometric Techniques in Fingerprint and Face Recognition, L C. Jain, U. Halici, I. Hayashi., S. B. Lee, and S. Tsutsui, eds., pp. 109-134, Boca Raton, Fla.: CRC Press, 1999.
- [24] C. L. Wilson, G. T. Candela, and C. I. Watson, "Neural Network Fingerprint Classification", J. Artificial Neural Networks, Vol. 1, No. 2, pp. 203-228, 1994.
- [25] L'O. Gorman and J. V. Nicherson, "An Approach to Fingerprint Filter Design", Pattern Recognition, Vol. 22, No. 1, pp.-29-38, 1989.
- [26] M. Kass and A. Witkin, "Analyzing Oriented Patterns", Computer Vision, Graphics, and Image Processing, Vol. 37, No. 3, pp.-362-385, March 1987.
- [27] A. K Jain, L. Hong, S. Pankanti, and R. Bolle, 'An Identity Authentication System Using Fingerprints', Proc. IEEE, Vol. 85, no. 9, pp1365-1388, Sept. 1997.
- [28] A. R. Rao and R. C. Jain, Computerized, "Flow Field Analysis: Oriented Texture Fields", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 14, no. 7., pp. 693- 709, July, 1992.
- [29] N. Ratha, S. Chen, and A. Jain, "Adaptive Flow Orientation Based Feature Extraction in Fingerprint Images", Pattern Recognition, vol. 28, pp. 1657-1672, Nov. 1995.
- [30] P. Perona, "Orientation Diffusions", IEEE Trans. Image Processing, vol.7, no.3, pp.457-467, Mar.1998.
- [31] Masahiro Kawagoe and Akio Tojo, "Fingerprint pattern classification", Pattern Recognition, Vol. 17, no.3, pp 295-303,1984 .
- [32] CHENG Xin-Ming, XU Dong-Cheng, XU Cheng, "A New Algorithm for Detecting Singular Points in Fingerprint Images", Third International Conference on Genetic and Evolutionary Computing, 2009.
- [33] Atiquzzaman, M., 1992. "Multiresolution Hough transform—an efficient method of detecting patterns in images", IEEE Trans. Pattern Anal.Machine Intell. 14 (11), 1090-1095.
- [34] Duda, R.O., Hart, P.E., 1972. "Use of Hough transformation to detect lines and curves in pictures", Commun. ACM 15 (1), 11-15.
- [35] Hough P.V.C., 1962. "Method and means for recognizing complex patterns", U.S. Patent No. 3069654.
- [36] D. S. Guru B, H. Shekar, P. Nagabhushan, "A simple and robust line detection algorithm based on small Eigen value analysis", Pattern Recognition. 25, pp. 1-13, 2003.

AUTHORS PROFILE

Praveen N received his MSc (Electronics) from Department of Electronics, Cochin University of Science and Technology, Kochi, 682022. He is working as Associate Professor of Electronics in the Department of Electronics, N S S College, Rajakumari, Idukki Dt., Kerala. He has 15 years of teaching experience and five years of research experience. At present he is doing PhD in Department of Electronics, Cochin University of Science and Technology. His research areas of interest include digital signal / image processing, biometrics, computer vision etc.

Dr.Tessamma Thomas received her M.Tech and Ph.D from Cochin University of Science and Technology, Cochin-22, India. At present she is working as Professor in the Department of Electronics, Cochin University of Science and Technology. She has to her credit more than 90 research papers, in various research fields, published in International and National journals and conferences. Her areas of interest include digital signal / image processing, bio medical image processing, super resolution, content based image retrieval, genomic signal processing etc.

Gender Effect Canonicalization for Bangla ASR

B.K.M. Mizanur Rahman Lecturer United International University, Dhaka, Bangladesh	Bulbul Ahamed Senior Lecturer Northern University Bangladesh, Dhaka, Bangladesh	Md. Asfak-Ur-Rahman Technical Project Manager Grype Solutions Bangladesh	Khaled Mahmud Lecturer Institute of Business Administration University of Dhaka	Mohammad Nurul Huda Associate Professor United International University, Dhaka, Bangladesh
--	---	---	--	---

Abstract—This paper presents a Bangla (widely used as Bengali) automatic speech recognition system (ASR) by suppressing gender effects. Gender characteristic plays an important role on the performance of ASR. If there is a suppression process that represses the decrease of differences in acoustic-likelihood among categories resulted from gender factors, a robust ASR system can be realized. In the proposed method, we have designed a new ASR incorporating the Local Features (LFs) instead of standard mel frequency cepstral coefficients (MFCCs) as an acoustic feature for Bangla by suppressing the gender effects, which embeds three HMM-based classifiers for corresponding male, female and gender-independent (GI) characteristics. In the experiments on Bangla speech database prepared by us, the proposed system has achieved a significant improvement of word correct rates (WCRs), word accuracies (WAs) and sentence correct rates (SCRs) in comparison with the method that incorporates Standard MFCCs.

Keywords- Acoustic Model; AutomaticSpeech Recognition; Gender Effects Suppression; Hidden Markov Model.

I. INTRODUCTION

The current automatic speech recognition (ASR) system had been investigated for achieving the adequate performance at any time and everywhere; however, it could not be able for providing the highest level accuracies till to date. One of the reasons is that the acoustic models (AMs) of a hidden Markov model (HMM)-based classifier include many hidden factors such as speaker-specific characteristics that include gender types and speaking styles. It is difficult to recognize speech affected by these factors, especially when an ASR system contains only a single acoustic model. One solution is to employ multiple acoustic models, one model for each type of gender. By handling these gender effects appropriately the robustness of each acoustic model in an ASR can be extended to some limit.

A method of decoding in parallel with multiple HMMs corresponding to hidden factors has recently been proposed in [1], [2] for resolving these difficulties. Multi-path acoustic modeling, that represents hidden factors with several paths in the same AM instead of applying multiple HMMs, was also presented [3]. Unfortunately, only a very few works have been done for ASR in Bangla (can also be termed as Bengali), which is one of the largely spoken languages in the world. More than 220 million people speak in Bangla as their native language. It is ranked sixth based on the number of speakers [4]. A major difficulty to research in Bangla ASR is the lack of proper speech corpus. Some efforts are made to develop

Bangla speech corpus to build a Bangla text to speech system [5].

Although some Bangla speech databases for the eastern area of India (West Bengal and Kolkata as its capital) were developed, but most of the natives of Bangla (more than two thirds) reside in Bangladesh, where it is the official language. Besides, the written characters of Standard Bangla in both the countries are same; there are some sounds that are produced variably in different pronunciations of Standard Bangla, in addition to the myriad of phonological variations in non-standard dialects [6]. Therefore, there is a need to do research on the main stream of Bangla, which is spoken in Bangladesh, ASR.

Some developments on Bangla speech processing or Bangla ASR can be found in [7]-[14]. For example, Bangla vowel characterization is done in [7]; isolated and continuous Bangla speech recognition on a small dataset using HMMs is described in [8].

Again, Bangla digit recognition was found in [15]. Since no work in Bangla was found for suppressing the gender factor, previously, we proposed a method [16] for that purpose by embedding multiple HMM-based classifiers. But this method of gender effect suppression did not incorporate any gender-independent (GI) classifier for resolving those male and female speakers whose voices have effect of opposite gender to some extent. To resolve this problem, we proposed another method for suppressing the gender factor more accurately by incorporating the GI classifier [17]; but this method could not be able to provide enough performance because of embedding standard mel frequency cepstral coefficients (MFCCs) as an input acoustic feature that did not include frequency domain information in its extraction process. Consequently, an exploitation of new feature is needed to obtain time and frequency domain information in its feature vector.

In this paper, we have designed a new ASR incorporating the Local Features (LFs) instead of standard mel frequency cepstral coefficients (MFCCs) as an acoustic feature for Bangla by suppressing the gender effects, which embeds three HMM-based classifiers for corresponding male, female and gender-independent (GI) characteristics. In the experiments on Bangla speech database prepared by us, the proposed system has achieved a significant improvement of word correct rates (WCRs), word accuracies (WAs) and sentence correct rates (SCRs) in

comparison with the method that incorporates Standard MFCCs.

This paper is organized as follows. Sections II discusses Bangla phoneme schemes, Bangla speech corpus and triphone model. On the other hand, Section III outlines MFCCs and LFs features extraction procedures and Section IV explain the gender effect suppression methods [16] and [17] including the proposed suppression technique by incorporating LFs. Section V describes the experimental setup and Section VI shows experimental results and provides a discussion. Finally, Section VII and VIII conclude the paper with some future remarks and references, respectively.

II. BANGLA PHONEME SCHEME, SPEECH CORPUS AND TRIPHONE MODEL

Bangla phonetic scheme and triphone models design were presented in our papers, [16] and [17]. These papers show ed how the left and right contexts are used to design the triphone models.

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

Hundred sentences from the Bengali newspaper “Prothom Alo” [18] are uttered by 30 male speakers of different regions of Bangladesh. These sentences (30x100) are used as male training corpus (D1). On the other hand, 3000 same sentences uttered by 30 female speakers are used as female training corpus (D2).

On the other hand, different 100 sentences from the same newspaper uttered by 10 different male speakers and by 10 different female speakers are used as male test corpus (D3) and female test corpus (D4), respectively. All of the speakers are Bangladeshi nationals and native speakers of Bangla. The age of the speakers ranges from 20 to 40 years. We have chosen the speakers from a wide area of Bangladesh: Dhaka (central region), Comilla – Noakhali (East region), Rajshahi (West region), Dinajpur – Rangpur (North-West region), Khulna (South-West region), Mymensingh and Sylhet (North-East region). Though all of them speak in standard Bangla, they are not free from their regional accent.

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

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

III. ACOUSTIC FEATURE EXTRACTOR

3.1 MFCC Feature Extractor

Conventional approach of ASR systems uses MFCC of 39 dimensions (12-MFCC, 12- Δ MFCC, 12- $\Delta\Delta$ MFCC, P, Δ P and $\Delta\Delta$ P, where P stands for raw energy of the input speech signal) and the procedure of MFCC feature extraction is shown in Fig. 1. Here, hamming window of 25 ms is used for extracting the feature. The value of pre-emphasis factor is 0.97.

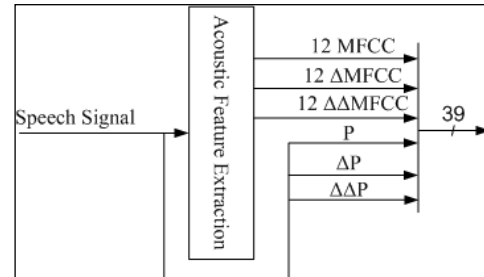


Fig. 1. MFCC feature extraction.

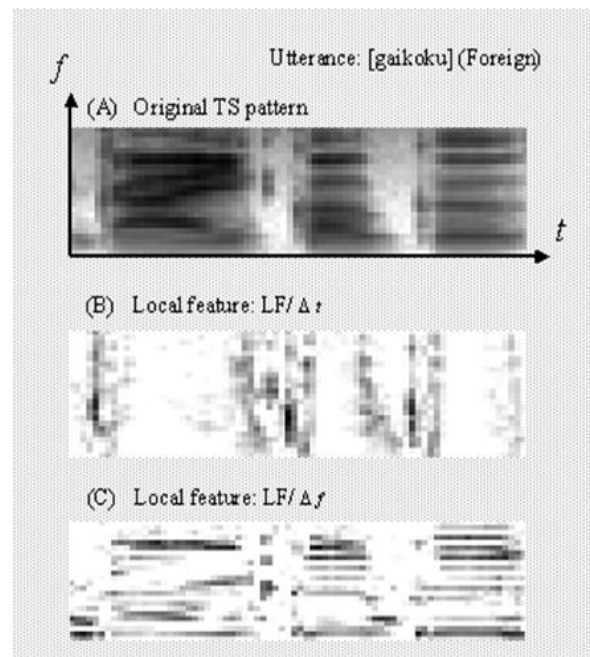


Fig. 2. Example of local features.

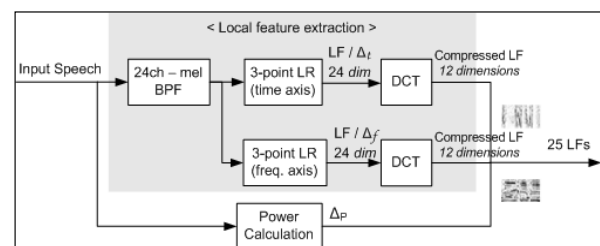


Fig. 3. Local feature extraction.

3.2 Local Feature Extractor

At the acoustic feature extraction stage, the input speech is first converted into LFs that represent a variation in spectrum along the time and frequency axes. Two LFs are then extracted by applying three-point linear regression (LR) along the time (t) and frequency (f) axes on a time spectrum pattern (TS), respectively. Fig. 2 exhibits an example of LFs for an input utterance, /gaikoku/. After compressing these two LFs with 24 dimensions into LFs with 12 dimensions using discrete cosine transform (DCT), a 25-dimensional (12 Δt , 12 Δf , and ΔP , where P stands for the log power of a raw speech signal) feature vector called LF is extracted. Fig.3 shows the local feature extraction procedure.

IV. GENDER EFFECT SUPPRESSION METHODS

4.1 MFCC-Based Suppression method without GI classifier [16]

Fig. 4 shows the system diagram of the existing method [16], where MFCC features are extracted from the speech signal using the MFCC extractor described in Section 3.1 and then male and female HMM classifiers are trained using the D1 and D2 data sets, respectively. Here, triphone acoustic HMMs are designed and trained using D1 and D2 data sets. Output hypothesis is selected based on maximum output probabilities after comparing male and female hypotheses, and passed the best matches hypothesis to the output.

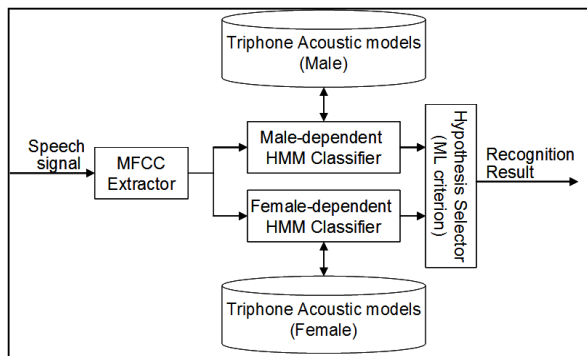


Fig. 4. MFCC-based gender effect suppression method without GI classifier [16].

4.2 MFCC-Based Suppression method incorporating GI classifier [17]

The diagram of the method [17] is depicted in Fig. 5. Here, the extracted MFCC features from the input speech signal are inserted into the male, female and GI HMM-based classifiers. The male, female and GI HMM-based classifiers are trained using the D1, D2 and (D1+D2) data sets. Here, output hypothesis is selected based on maximum output probabilities after comparing male, female and gender independent hypotheses, and passed the best matches hypothesis to the output.

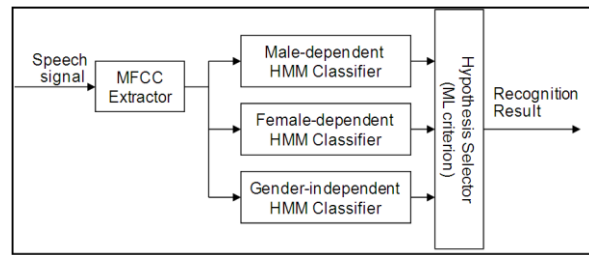


Fig. 5. MFCC-based gender effect suppression method incorporating GI classifier [17].

4.3 LF-based Proposed Suppression Method incorporating GI classifier

The diagram of the LF-based method is depicted in Fig. 6. Here, the extracted LFs [19] from the input speech signal are inserted into the male, female and GI HMM-based classifiers. The male, female and GI HMM-based classifiers are trained using the D1, D2 and (D1+D2) data sets. Here, output hypothesis is selected based on maximum output probabilities after comparing male, female and gender independent hypotheses, and passed the best matches hypothesis to the output.

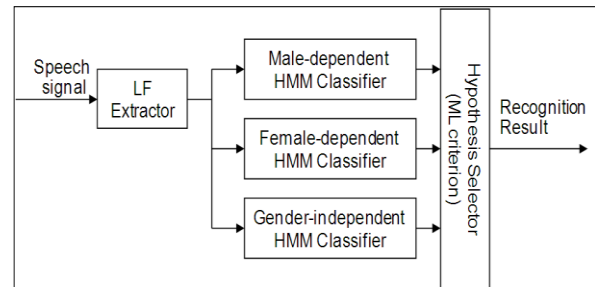


Fig. 6. Proposed LF-based suppression method incorporating GI classifier.

V. EXPERIMENTAL SETUP

The frame length and frame rate are set to 25 ms and 10 ms (frame shift between two consecutive frames), respectively, to obtain acoustic features (MFCCs and LFs) from an input speech. MFCC and comprised of 39 (12-MFCC, 12- Δ MFCC, 12- $\Delta\Delta$ MFCC, P, ΔP and $\Delta\Delta P$, where P stands for raw energy of the input speech signal) and 25 (12 delta coefficients along time axis, 12 delta coefficients along frequency axis, and delta coefficient of log power of a raw speech signal) dimensions, respectively.

For designing an accurate continuous word recognizer, word correct rate (WCR), word accuracy (WA) and sentence correct rate (SCR) for (D3+D4) data set are evaluated using an HMM-based classifier. The D1 and D2 data sets are used to design Bangla triphones HMMs with five states, three loops, and left-to-right models. Input features for the classifier are 39 dimensional MFCCs and 25 dimensional LFs.

In the HMMs, the output probabilities are represented in the form of Gaussian mixtures, and diagonal matrices are used. The mixture components are set to 1, 2, 4 and 8.

For evaluating the performance of the methods, [16] and [17], we have designed the following experiments:

- (a) MFCC (Train: 3000 male, Test: 1000 male+1000 female).
- (b) MFCC (Train: 3000 female, Test: 1000 male+1000 female).
- (c) MFCC (Train: 3000 male + 3000 female, Test: 1000 male +1000 female).
- (d) MFCC (Train: 3000 male, Train: 3000 female, Test: 1000 male + 1000 female) [16].
- (e) MFCC (Train: 3000 male, Train: 3000 female, Train: 3000 male + 3000 female, Test: 1000 male + 1000 female).

New experiments given below are designed for mixture component one using the LFs for evaluating the performance:

- (a) LF (Train: 3000 male, Test: 1000 male+1000 female).
- (b) LF(Train: 3000 female, Test: 1000 male+1000 female).
- (c) LF(Train: 3000 male + 3000 female, Test: 1000 male +1000 female).
- (d) **Proposed**, LF(Train: 3000 male, Train: 3000 female, Train: 3000 male + 3000 female, Test: 1000 male + 1000 female).

VI. EXPERIMENTAL RESULTS AND ANALYSIS

Fig. 7 shows the comparison of word correct rates among all the MFCC-based investigated methods, (a), (b), (c), (d) and (e). Among all the mixture components investigated, the method, (e) shows higher performance in comparison with the other method evaluated. It is noted that the method, (e) exhibits its best performance (92.17%) at mixture component two.

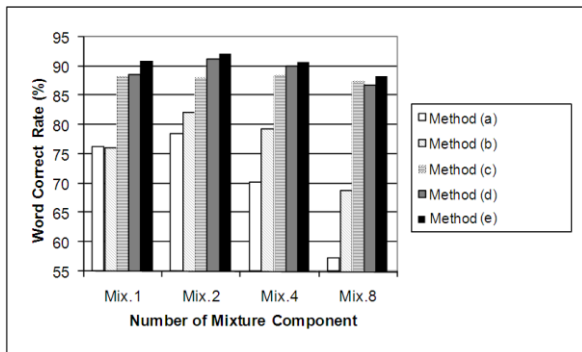


Fig. 7. Word correct rate comparison among MFCC-based investigated methods.

Word accuracies for different investigated methods based on MFCCs are depicted in Fig. 8. From the figure, it is observed that the highest level performance (91.64%) at mixture component two is found by the method, (e) compared to the other methods investigated. Here, the performance of the methods, (a), (b), (c), (d) and (e) at mixture component two

are 77.58%, 81.47%, 87.39%, 90.78% and 91.64%, respectively.

It is shown from the Fig. 9 that sentence correct rates for the MFCC-based investigated methods, (a), (b), (c), (d) and (e) are 77.20%, 81.45%, 86.60%, 90.45% and 91.30%, respectively, where the method, (e) provides its best performance. The methods, (a), (b) and (c) give less performance in comparison with the method (d) because (d) incorporates both HMM-based classifiers for male and female. Again, the method, (e) incorporates GI HMM-based classifier over the method (d), which increases sentence correct rate significantly. Since the maximum output probability is generated by the MFCC based method, (e) after comparing the probabilities among male, female and GI classifiers, the suppression method, (e) shows its superiority.

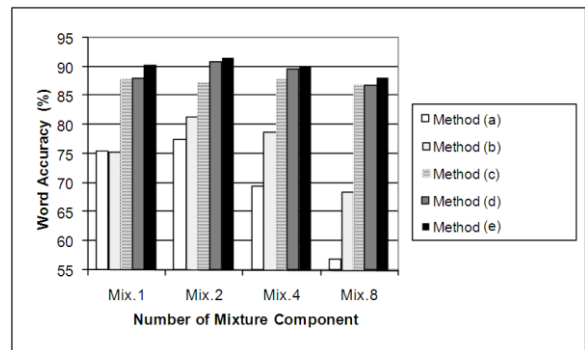


Fig. 8. Word accuracy comparison among MFCC-based investigated methods.

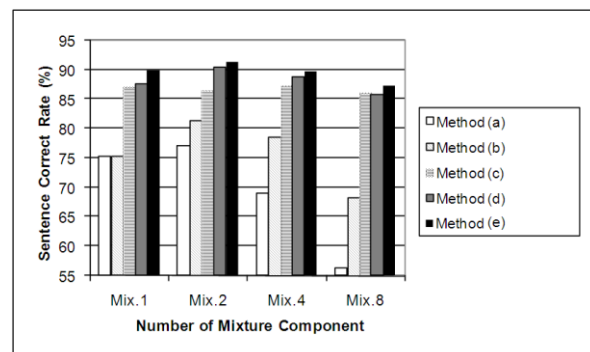


Fig. 9. Sentence correct rate comparison among MFCC-based investigated methods.

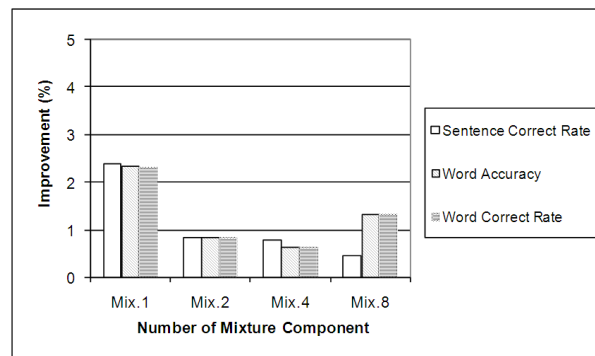


Fig. 10. Effect of Gender-Independent classifier in the MFCC-based method (e) over the method (d).

Improvements by the GI classifier in the MFCC-based method (e) over the method (d) that does not incorporate GI classifier is shown in Fig. 10. From the figure, it is observed that the method, (e) shows its highest level improvement at mixture components one, where the improvement of sentence correct rate, word accuracy and word correct rate are 2.4%, 2.36% and 2.32%, respectively.

Word Correct Rates (WCRs), Word Accuracies (WAs) and Sentence Correct Rates (SCRs) for LF and MFCC based methods, (a), (b), (c) and (d)/(e) using (D3+D4) data set are shown in Table I. Here, methods (d) and (e) represent LF-based and MFCC-based suppression methods with gender independent (GI) classifiers, respectively. From the experiment #1, where the HMM-based classifier is trained with D1 data set and evaluated with (D3+D4) data set, a tremendous improvement of WCRs, WAs and SCRs are exhibited by the LF-based ASR that incorporates gender effect canonicalization module in the ASR process. Similarly, the same pattern of performance is also achieved in the experiment #2, where the HMM-based classifier is trained with D2 data set and evaluated with (D3+D4) data set. This trend explicates the LFs, which embeds time and frequency domain information in its extraction procedure, as excellent feature for the Bangla automatic speech recognition system. Again, the experiment #3, which shows the GI ASR for Bangla language based on LF and MFCC features, is trained with (D1+D2) and evaluated with (D3+D4) data sets and provides 3.83%, 2.60% and 3.70% improvements by the LF-based method in comparison with the MFCC-based counterpart. Finally, the methods in the experiment #4 imply two ASR systems for GI Bangla ASR by integrating gender factor canonicalization process in its architecture and shows the highest level performance for WCRs, WAs and SCRs compare to the corresponding the methods in the experiments #1, #2 and #3. Since these methods of the experiment #4 always maximize the output probabilities obtained from the three classifiers: male, female and GI, it shows its maximum level of performance. Besides, the LF-based methods improves the WCRs, WAs and SCRs by 3.94%, 3.43% and 3.90%, respectively than the method that inputs MFCC features in the HMM-based classifier in the gender effect suppression process.

Table I. Word Correct Rates (WCRs), Word Accuracies (WA) and Sentence Correct Rates (SCRs) for LF and MFCC based methods, (a), (b), (c) and (d)/(e) using (D3+D4) data set for mixture component one. Here, methods (d) and (e) are LF-based and MFCC-based suppression methods with gender independent (GI) classifiers, respectively.

Experiments	Methods	WCR(%)	WA(%)	SCR(%)
#1	LF-based (a)	88.63	85.53	86.60
	MFCC-based (a)	76.38	75.41	75.30
#2	LF-based (b)	86.29	83.65	84.50
	MFCC-based (b)	76.16	75.50	75.45
#3	LF-based (c)	92.22	90.43	90.85
	MFCC-based (c)	88.39	87.83	87.15
#4	LF-based (d)	94.94	93.86	93.95
	MFCC-based (e)	91.00	90.43	90.05

On the other hand, Table II exhibits sentence recognition performance for LF and MFCC based methods, (a), (b), (c) and (d)/(e) using (D3+D4) data set, where the methods (d) and (e) represent LF-based and MFCC-based suppression methods with gender independent (GI) classifiers, respectively. Here, experiments #1, #2, #3 and #4 use same corpora for training and evaluation that we explained earlier. From all the experiments, it is evident that the LF-based method reduces the number of incorrectly recognized sentences with respect to its counterpart. The experiment #4 shows the highest number of correctly recognized sentences than the corresponding methods in the other experiments, #1, #2 and #3 investigated. For an example, in the LF-based and MFCC-based methods of experiment #4, the numbers of correctly recognized sentences are 1879 and 1801, respectively that are the highest numerical figures among the corresponding methods of all the experiments. It is noted that two significant phenomenon contributed more for obtaining the best experimental results by the LF-based method of experiment #4: i) both time and frequency domain information and ii) selection of maximum probability among the three output probabilities calculated by the male, female and GI HMM-based classifiers.

Table II. Sentence recognition performance for LF and MFCC based methods, (a), (b), (c) and (d)/(e) using (D3+D4) data set for mixture component one. Here, methods (d) and (e) are LF-based and MFCC-based suppression methods with gender independent (GI) classifiers, respectively

Experiments	Methods	Sentence Recognition Performance (Total 2000 Sentences)	
		Correctly Recognized Sentences, H	Incorrectly Recognized Sentences, S
#1	LF-based (a)	1732	268
	MFCC-based (a)	1506	494
#2	LF-based (b)	1690	310
	MFCC-based (b)	1509	491
#3	LF-based (c)	1817	183
	MFCC-based (c)	1743	257
#4	LF-based (d)	1879	121
	MFCC-based (e)	1801	199

Moreover, word recognition performance for LF and MFCC based methods, (a), (b), (c) and (d)/(e) using (D3+D4) data set are summarized in the Table III. In the table, methods (d) and (e) represent LF-based and MFCC-based suppression methods with gender independent (GI) classifiers, respectively. The same speech corpora for training and evaluation are used for the experiments #1, #2, #3 and #4 which is already described in the earlier. It is observed from all the experiments that the LF-based method increases the number of correctly recognized words in comparison with its counterpart. The highest number of correctly recognized words shown by the experiment #4 with respect to their corresponding methods in the other experiments, #1, #2 and #3 are evident from the table. It can be mentioned as an example that the LF-based and MFCC-based methods of experiment #4 provide the highest numbers of correctly recognized words, which are 6247 and 5988, respectively, dictates the respective numerical figures obtained for the corresponding methods of all the other experiments. The reason for obtaining the best experimental results by the LF-based method of experiment #4 is also illustrated earlier.

Table III. Word recognition performance for LF and MFCC based methods, (a), (b), (c) and (d)/(e) using (D3+D4) data set for mixture component one. Here, methods (d) and (e) represent LF-based and MFCC-based suppression methods with gender independent (GI) classifiers, respectively.

Experiments	Methods	Word Recognition Performance (Total 6580 words)			
		Correctly Recognized Words, H	Deletion, D	Substitution, S	Insertion, I
#1	LF-based (a)	5832	58	690	204
	MFCC-based (a)	5026	438	1116	64
#2	LF-based (b)	5678	89	813	174
	MFCC-based (b)	5011	448	1121	43
#3	LF-based (c)	6068	55	457	118
	MFCC-based (c)	5816	152	612	37
#4	LF-based (d)	6247	33	300	71
	MFCC-based (e)	5988	140	452	38

VII. CONCLUSION

This paper has proposed an automatic speech recognition technique based on LFs for Bangla language by suppressing the gender effect incorporating HMM-based classifiers for male, female and Gender-Independent characteristics. The following information concludes the paper.

- i) The MFCC-based method incorporating GI classifier provides the higher performance than the method that does not incorporate GI classifier. The MFCC-based method incorporating GI exhibits its superiority at all the mixture components investigated.
- ii) The incorporation of GI HMM classifier improves the word correct rates, word accuracies and sentence correct rates significantly.
- iii) The proposed LF-based method shows a significant improvement of word correct rates, word accuracies and sentence correct rates for mixture component one.

In future, the authors would like to evaluate performance by incorporating neural network based systems.

REFERENCES

- [1] S. Matsuda, T. Jitsuhiro, K. Markov and S. Nakamura, "Speech Recognition system Robust to Noise and Speaking Styles," Proc. ICSLP'04, Vol.IV, pp.2817-2820, Oct. 2004.
- [2] T. Shinozaki and S. Furui, "Spontaneous Speech Recognition Using a Massively Parallel Decoder," Proc. ICSLP'04, Vol.III, pp.2817-2820, Oct. 2004.
- [3] A. Lee, Y. Mera, K. Shikano and H. Saruwatari, "Selective multi-path acoustic model based on database likelihoods," Proc. ICSLP'02, Vol.IV, pp.2661-2664, Sep. 2002.
- [4] http://en.wikipedia.org/wiki/List_of_languages_by_total_speakers, Last accessed July04, 2012.
- [5] S. P. Kishore, A. W. Black, R. Kumar, and Rajeev Sangal, "Experiments with unit selection speech databases for Indian languages," Carnegie Mellon University.
- [6] S. A. Hossain, M. L. Rahman, and F. Ahmed, "Bangla vowel characterization based on analysis by synthesis," Proc. WASET, vol. 20, pp. 327-330, April 2007.
- [7] A. K. M. M. Houque, "Bengali segmented speech recognition system," Undergraduate thesis, BRAC University, Bangladesh, May 2006.
- [8] K. Roy, D. Das, and M. G. Ali, "Development of the speech recognition system using artificial neural network," in Proc. 5th International

Conference on Computer and Information Technology (ICCIT02), Dhaka, Bangladesh, 2003.

- [9] M. R. Hassan, B. Nath, and M. A. Bhuiyan, "Bengali phoneme recognition: a new approach," in Proc. 6th International Conference on Computer and Information Technology (ICCIT03), Dhaka, Bangladesh, 2003.
- [10] K. J. Rahman, M. A. Hossain, D. Das, T. Islam, and M. G. Ali, "Continuous bangla speech recognition system," in Proc. 6th International Conference on Computer and Information Technology (ICCIT03), Dhaka, Bangladesh, 2003.
- [11] S. A. Hossain, M. L. Rahman, F. Ahmed, and M. Dewan, "Bangla speech synthesis, analysis, and recognition: an overview," in Proc. NCCPB, Dhaka, 2004.
- [12] S. Young, et al, The HTK Book (for HTK Version. 3.3), Cambridge University Engineering Department, 2005. <http://htk.eng.cam.ac.uk/prot-doc/ktkbook.pdf>.
- [13] http://en.wikipedia.org/wiki/Bengali_script, Last accessed September 28, 2011.
- [14] C. Masica, *The Indo-Aryan Languages*, Cambridge University Press, 1991.
- [15] Ghulam Muhammad, Yousef A. Alotaibi, and Mohammad Nurul Huda, "Automatic Speech Recognition for Bangla Digits," ICCIT'09, Dhaka, Bangladesh, December 2009.
- [16] Mohammed Rokibul Alam Kotwal, Foyzul Hasan and Mohammad Nurul Huda, "Gender effects suppression in Bangla ASR by designing multiple HMM based classifiers," CICN 2011, Gwalior, India.
- [17] Foyzul Hasan, Mohammed Rokibul Alam Kotwal and Mohammad Nurul Huda, "Bangla ASR design by suppressing gender factor with gender-independent and gender-based HMM classifiers," WICT 2011, December 2011, Mumbai, India.
- [18] Prothom Alo. Online: www.prothom-alo.com.
- [19] T. Nitta, "Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA," Proc. ICASSP'99, pp.421-424, 1999.

AUTHORS PROFILE

B.K.M. Mizanur Rahman was born in Jhenaidah, Bangladesh in 1972. He completed his B.Sc. in Electrical and Electronic Engineering Degree from BUET, Dhaka, Bangladesh. He is a student of Masters in Computer Science and Engineering at United International University, Dhaka, Bangladesh. He is working as a Lecturer in the Department of Electrical and Electronic Engineering of the same university. His research interests include Speech Recognition, Digital Signal Processing and Renewable Energy.

Bulbul Ahamed was born in Munshiganj, Bangladesh in 1982. He obtained his B. Sc. in Computer Science and Engineering and MBA (major in MIS & Marketing) from Northern University Bangladesh. Now he is pursuing his M.Sc. in Computer Science and Engineering at United International University, Bangladesh. He is now working as Senior Lecturer in Northern University Bangladesh. His research interests include Speech Recognition, Artificial Intelligence, Neural Network and Business. He has published his articles in different journals and conferences of USA, Pakistan, United Arab Emirates and Bangladesh.

Md.Asfak-Ur-Rahman was born in Kishoreganj, Bangladesh in 1985. He obtained his B. Sc. in Computer Science and Engineering BRAC University, Bangladesh. Now he is pursuing his M.Sc. in Computer Science and Engineering at United International University, Bangladesh. He is now working as Technical Project Manager in Grype Solutions, Bangladesh. His research interests include Speech Recognition, Artificial Intelligence and web programming.

Khaled Mahmud was born in 1984 at Pabna, Bangladesh. He was graduated from Bangladesh University of Engineering and Technology (BUET) in Computer Science and Engineering. He had his MBA (Marketing) from Institute of Business Administration, University of Dhaka. He was awarded gold medals both in his secondary and higher secondary school level for excellent academic performance. He is a Fulbright Scholar, now pursuing his MBA at Bentley University, Massachusetts, USA. He previously worked as Assistant Manager in Standard Chartered Bank. He has research interest business, technology, e-learning, e-governance, human resource management and social issues. He has published his articles in

journals and conferences of USA, Canada, Australia, United Arab Emirates, Malaysia, Thailand, South Korea, India and Bangladesh.

Mohammad Nurul Huda was born in Lakshmipur, Bangladesh in 1973. He received his B. Sc. and M. Sc. in Computer Science and Engineering degrees from Bangladesh University of Engineering & Technology (BUET), Dhaka in 1997 and 2004, respectively. He also completed his Ph. D from the

Department of Electronics and Information Engineering, Toyohashi University of Technology, Aichi, Japan. Now, he is working as an Associate Professor in United International University, Dhaka, Bangladesh. His research fields include Phonetics, Automatic Speech Recognition, Neural Networks, Artificial Intelligence and Algorithms. He is a member of International Speech Communication Association (ISCA).

Three Layer Hierarchical Model for Chord

Waqas A. Imtiaz, Shimul Shil, A.K.M Mahfuzur Rahman

Abstract— Increasing popularity of decentralized Peer-to-Peer (P2P) architecture emphasizes on the need to come across an overlay structure that can provide efficient content discovery mechanism, accommodate high churn rate and adapt to failures in the presence of heterogeneity among the peers. Traditional p2p systems incorporate distributed client-server communication, which finds the peer efficiently that store a desired data item, with minimum delay and reduced overhead. However traditional models are not able to solve the problems relating scalability and high churn rates. Hierarchical model were introduced to provide better fault isolation, effective bandwidth utilization, a superior adaptation to the underlying physical network and a reduction of the lookup path length as additional advantages. It is more efficient and easier to manage than traditional p2p networks. This paper discusses a further step in p2p hierarchy via 3-layers hierarchical model with distributed database architecture in different layer, each of which is connected through its root. The peers are divided into three categories according to their physical stability and strength. They are Ultra Super-peer, Super-peer and Ordinary Peer and we assign these peers to first, second and third level of hierarchy respectively. Peers in a group in lower layer have their own local database which hold as associated super-peer in middle layer and access the database among the peers through user queries. In our 3-layer hierarchical model for DHT algorithms, we used an advanced Chord algorithm with optimized finger table which can remove the redundant entry in the finger table in upper layer that influences the system to reduce the lookup latency. Our research work finally resulted that our model really provides faster search since the network lookup latency is decreased by reducing the number of hops. The peers in such network then can contribute with improve functionality and can perform well in P2P networks.

Keywords- Hierarchy; DHT; CHORD; P2P.

I. INTRODUCTION

Peer-to-Peer (P2P) network is a logical overlay network which is built on top of one or more existing physical networks. P2P networks, over the last two decades has been recognized as a more efficient and flexible approach for sharing resources, compared to the traditional Client-Server model. Internet-scale decentralized architecture bases on p2p, created an environment for millions of users, allowing them to simultaneously connect and share content with ease and reliability [1][8]. Efficient data location, lookups, redundant storage and distributed content placement of p2p overlay networks have raised a big deal of attention, not only for researchers/academicians but also in practical usage of the technology [7]. The distributed approach of p2p system is less vulnerable to attacks, robust and highly available as compared to its client-server counterpart.

P2P algorithms are a class of decentralized distributed systems collectively, called as Distributed Hash Tables (DHTs). DHT is a distributed data structure whose main

function is to hold the key-value pair in a completely distributed manner and any participating peer in the overlay network can retrieve the value associated with the given key [8]. DHT uses consistent hashing to map the responsible node for a key-value pair. Along with efficient mapping of a key-value pair to nodes, DHT also has the ability to isolate the network changes to a small part of it thus limiting the overhead and bandwidth consumption during networks and resources updates [8]. DHT is an abstract idea that helps us to achieve complete independence from a central lookup entity and tolerance to changes in the network [9]. Different algorithms have been designed to implement and refine DHT idea. CAN, PASTRY, TAPSTERY, CHORD are the most popular implementations of DHT. We choose chord because of well-organized data structures and efficient routing schemes.

Chord is an efficient distributed lookup service based on consistent hashing which provides support for only one operation: given a key and it efficiently maps the key onto a node [7]. At its heart chord provides a fast and efficient computation of the hash function by mapping keys onto the nodes [7]. Chord basically creates a one dimensional identifier circle which ranges from 0 to $(2^m - 1)$, where m is the number of bits on the identifier circle and every node and key on the identifier circle are assigned an m -bit identifier. Node identifier is obtained by hashing the port number and node's IP address, whereas key identifier, 160 bits, is created by hashing the key [7]. Identifiers are assigned in such a way that the probability of two nodes having same hashed identifiers is negligible [7].

The main advantage of Chord in a large-scale setting is its high scalability and self-organization. But with the large amount of users, high churn condition, introduce a high overhead while maintaining the DHT structure. Further, dynamic joining and leaving of nodes may results in loss of key-value pairs, even though their respective nodes are still alive. Controlling this situation is difficult and incurs more overhead. While majority of the peers are short lived and have minimal capabilities, a small percentage typically remains up for long periods and have relatively better storage, bandwidth and memory. This property has been used to design hierarchical models, where more stable nodes can dynamically form an upper level overlay. Other short lived peers can get themselves connected as sub-overlay of these upper level nodes.

Naturally hierarchical DHT design has a better fault isolation, effective bandwidth utilization, superior adaptation to the underlying physical network and a reduction of the lookup path length as an additional advantage [14]. It is more efficient and easier to manage than the pure p2p structure. Moreover by dividing the whole system into several different layers that tries to solve local tasks inside their own layers, results in reduced workload and efficient operation of the network [48].

Keeping in view the advantages of hierarchical models over traditional p2p models, this paper proposes a three layer DHT based overlay, with decentralized key database architecture that relies on p2p algorithm Chord. This projected three layer p2p architecture with decentralized database consists of a number of database subsystems, each of which is connected to others through its root only.

To meet the challenges of present internet applications, this paper presents a three hierarchical model presented in [1]. The lower layer peers are designed to be deployed on resource constraints. In p2p overlay network lower layer peers do not need to deliver high data rates and high computational power. The second layer act as a medium between lower and upper layer. Middle layer is designed by the super peer. This super peer performs as a central server for lower layer peers. The upper layer is the core or backbone layer in this architecture. We design this layer with ring based ultra-superpeer communication. Ultra-superpeer in upper layer also acts as a central server for associated middle layer super-peers. Finally, each layer super-peers perform as a central server for immediate lower layer peers. Each ultra-superpeer maintains the pointer index and a finger table that contains the successor/predecessor list in chord ring and index contains the all previous level peer ID and data list, but does not store the data or document. Moreover, super-peer in the middle layer only maintains the pointer index. Each of the ordinary peers is attached to a super peer with point-to-point or mesh topology connection and this super peer belongs to both the ordinary peers and the super-peers. As a result local peers do not have to share the burden of possibly high maintenance traffic and the overlay network does not have to deal with their performance bottlenecks and low reliability [7].

II. SYSTEM MODEL

Based on the capacity and availability of peers in the overlay network, they are classified into three categories: Ordinary peers, Super-peers and Ultra Super-peers as shown in fig. 1. Each peer in the overlay network has a peer ID, which is computed from its IP address and port address pair using consistent hash function (SHA-1) [10]. Distributed data sharing techniques are applied in the hierarchical p2p model. A data item is represented by a key-value pair. Where key is the data name and value is the content associated with the key [10]. A global consistent hash function is used to map object key to peer identifiers. A peer uses a *put (key, value)* command to insert the data item and *get (key)* command for obtaining the value of the data item in the overlay [10]. In the 3-layer hierarchical based p2p system, if the lower layer ordinary peer wants to share the file, it is necessary to send the metadata information to its associated superpeer in the middle layer.

The superpeer keeps the metadata in the index. As shown in fig. 2, superpeer index has two field, key_ID and peer_ID. Where key_ID is the file sharing identity and peer_ID is the ordinary peer identity. However, it is necessary to further declare the metadata information to the associated ultra-superpeer in extended format that are kept in the index of the ultra-superpeer. The extended index has three field, key_ID, peer_ID and superpeer_ID as shown in fig. 3. The

superpeer_ID is the associated superpeer of the ordinary peer. At the query process, the metadata index is to provide the sufficient information [15].

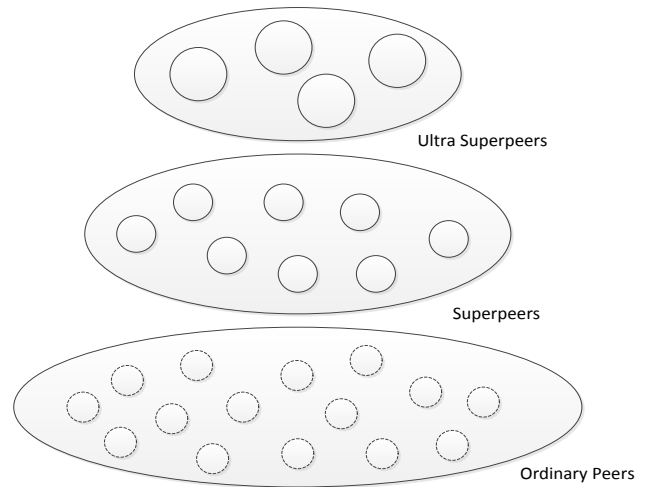


Fig 1: Three layer hierarchical Model for Chord

Key_ID	Peer_ID
X	N12
.....
.....

Key_ID	Peer_ID	Superpeer_ID
X	N12	N1
.....
.....
.....

Fig 3: Ultra Superpeer Index

A. Ordinary Peers

The ordinary peers form the single connection structure where peers communicate only with their super-peers [14]. All peers frequently join and leave the lower layer without affecting the entire overlay network. Ordinary peers are the members of the lower layer hierarchy and may disconnect after sharing their resources over p2p network. Ordinary peer can communicate with other ordinary peers through associated super-peers in the second layer.

In the single-connection structure, every ordinary peer uses the PING/PONG algorithm periodically to check the connectivity with super-peer. The ordinary peers send the PING message to its super-peer and the super-peer responds with a PONG message [7] [14]. As a result, at least two hops are needed to communicate with each other [13] [14].

Ordinary peers become a super-peer by sending an upgrade (capability and availability) level request to the associated super-peers. The current super-peer creates a list of ordinary peer request which may become a super-peer after leaving the current super-peer [11]. Ordinary peers ID in the lower layer groups must be smaller than the associated super-peer ID.

B. Super Peers

Peers with more stability and storage are placed in the middle layer of hierarchical system. Super-peers are also organized by the single connection structure. So that, middle layer super-peers check their connectivity with the upper layer via PING/ PONG message. Moreover, super-peers act as a server to maintain the index pointer information of ordinary peers and are also responsible for the query an object in the overlay network [7] [13].

In our hierarchical architecture, peers ID in the lower layer groups must be smaller than the associated super-peer ID and the super-peer ID in the middle layer must lay between the ultra-superpeer ID and predecessor ID in the upper layer. Every super-peer contains the metadata of the lower layer ordinary peers as index pointer.

C. Ultra Super peers

Ultra super-peer are peers with ideal characteristic, higher availability and which are being predicted to be available in future for a long period compared to the middle layer super-peer of the hierarchical structure. Ultra super-peers are organized in a circle where peers are connected to one another similar to that of a chord network. Moreover, ultra super-peer is the gateway or path to communicate among the super-peers in the middle layer, and acts as a central server for associated middle layer super-peers. In the chord ring each ultra-superpeers apply the *stabilization* protocol periodically to update the successor / predecessor pointer and the optimized finger table in the event of peer failure or migration from middle layer [7] [13] [14].

III. NODES PLACEMENT

Proposed three layer hierarchical model, applies different methods for joining and leaving of peers in overlay network.

A. Peers Joining

Peer that joins that overlay network, will become an ordinary peer. Ordinary peer cannot directly opt of becoming a superpeer or an ultra-superpeer. Ordinary peers can only migrate to the immediate upper layer based on their physical strengths i.e. bandwidth, stability, storage space etc.

The joining peer using any bootstrapping method sends a joining request to an existing known superpeer. Superpeer examines the peer ID of the joining peer and finds out a place for the joining peer in the overlay network [7]. After finding a place, superpeer sends a response message to the joining peer with the information of joining place. Ordinary peer on receiving the necessary information joins the overlay network and sends the metadata to its associated superpeer. Corresponding ultra superpeer is also informed of the new entry by its associated superpeer [11] [12].

B. Peers Leaving

Peers are classified into three layers that comprise the hierarchical model. Peers from each layer can leave the overlay network in the following manner:

1) *Ordinary Peer Leaves:* Ordinary peers may leave the hierarchy without informing any other peers because it does not

maintain the routing table. Superpeer detects the leaving of an ordinary peer by PING/PONG message in the lower layer. When superpeer detects a leaving ordinary peer, it changes the status of the leaving peer by deleting its associated metadata. Superpeer also informs the associated ultra superpeer in upper layer to delete the metadata of the corresponding ordinary peer.

2) *Super Peer Leaves:* As superpeer is responsible for multiple tasks in the hierarchy, so it cannot leave the system without any prior arrangements. Before superpeer leaves the system, it must select a candidate superpeer from the ordinary peers in the group.

After selecting a new superpeer, all metadata is transferred to the new superpeer. The new superpeer migrates to the middle layer and operates in place of the leaving superpeer. Migrated candidate superpeer also informs its connected ultra-superpeer to update the metadata and its associated ordinary peers about its arrival.

3) *Ultra Super Peer Leaves:* Ultra-superpeer selection and migration in the overlay network is similar to the superpeers leaving process, except the backup of metadata. Ultra-superpeers use the DHT chord protocol to periodically updates itself and uses its successor from finger table as backup of the metadata. The new ultra-superpeer on joining the upper layers, also informs its successor. When the candidate ultra-superpeer migrates to the upper layer, at the same time candidate superpeer also migrates to the middle layer to fix the link among the three layers. It is exceptional to leaves the ultra superpeer because of its stability.

C. Peers Migration

When a superpeer in the middle layer fails or leaves the overlay network, a new superpeer will be needed to establish communication between the associated ordinary peers and to perform the query process [7]. Ordinary peers previously marks as candidate superpeers are selected to replace the failed superpeer [1]. If the system cannot find any candidate superpeer, then an ordinary peer is selected on the basis of capacity, firewall support and availability. After finding the candidate Superpeer, it is migrated to the middle layer in place of the leaving Superpeer. Same procedure is also applied between the middle and upper layer in hierarchy, if the ultra superpeer fails or leaves. As a result, the number of Superpeers in the middle layer and ultra-superpeers in the upper layer is possibly unchanged.

IV. LOOKUP PROCESS

Lookup process of the proposed 3 layer hierarchy is shown in fig. 4. Lookup process starts at the ordinary peer, where data ID of a particular data item is needed to look up the peers. The data ID is obtained by hashing the data key [10]. Query message containing the required data ID is forwarded from the ordinary peer to the associated superpeer. After getting the query request, superpeer checks its database against the provided data ID. If the data ID is found in the ordinary peer groups, lookup process is done, otherwise the query request is forwarded to the connected ultra-superpeer in the ring [1][14]. When the query request arrives at associated ultra- superpeer,

it checks its own database. If data item is not found, it forwards the query to its successor based on the finger table similar to chord ring, and tries to find the data item. Ultra-superpeers in upper layer collectively contain all the overlay peers metadata information. So that data item is found efficiently, if it is available in the overlay.

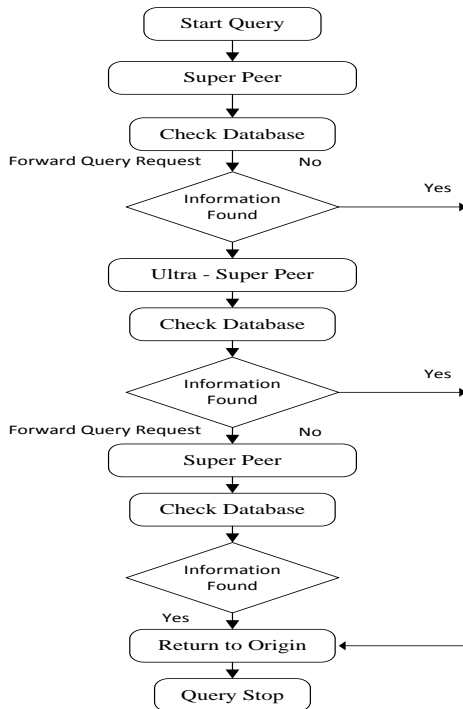


Fig 4: Hierarchical Model lookup process

V. ANALYTICAL ANALYSIS

Performance of the proposed system can be measured on the basis of lookup latency and hoops per lookup. We compare conventional chord with our proposed three layer hierarchical chord to identify the efficient model for p2p applications.

For conventional chord we have assumed the following parameters [11][14].

- Peer and key identifiers are random.
- Number of peers = N
- Number of entries in Finger table = $O(\log N)$
- Probability of the number of forwarding = $O(\log N)$

Because of the stable finger table in conventional chord, the average hops required to complete a lookup process is $\frac{1}{2} \log_2 N$ [1], where N is the number of peers in the overlay network. For our proposed three layer hierarchical model, we assume single connection structures, and take the following parameters for comparison [7][11][14][15]:

- Total number of peers in the overlay = N
- Number of ultra-superpeers = U
- Number of superpeers = S
- Number of Ordinary peers = $\{N - (S + U)\}$

- Number of peers in a lower layer group = $\frac{N-(S+U)}{s}$
- Probability of finding data item at super node = Q
- Number of entries in Finger table = $O(\log U)$
- Time complexity is $O(\log U + 1 + 1)$

Time complexity is the least amount of time required to execute a process or program. When the number of peers in the overlay network is N , the typical time complexity of searching is $O(N)$ for unstructured p2p network and $O(\log N)$ for structured p2p network [1]. We have used unstructured topology in the lower two layers and structure topology in the upper layer. In lower and middle layer, we apply the single connection intra group structure, therefore the typical time complexity of searching is $O(1)$, as single peer is traversed in order to complete the query the query process. In upper layer, the typical time complexity of searching is $O(\log U)$ because we apply structured based chord protocol. Finally, total time complexity of the model is $O(\log U + 1 + 1)$.

In our proposed hierarchy model, three cases occur during lookup process i.e. average number of hops required 1, if the query process is done between ordinary peers and superpeers. Average number of hops required is $(1+1) = 2$, if the query process is done among ordinary peers, super peers and associated ultra-superpeer (without using chord ring) because of single connection structure. Average number of hops required $\frac{1}{2} \log U$, if the query process is done in upper layer chord ring. Thus total number of hops required to perform a query process in $\frac{1}{2} \log (U+2)$.

Lookup latency considered in our model is half of the time required for a peer in the traditional chord. Because in the hierarchical model, ultra-superpeers and superpeers, performs the lookup process and have more physical capability then the ordinary peers. That is why we assume the average link latency of a peer in our hierarchical system is 50 % less than the traditional chord peer which are unstable and have minimum computation capabilities.

To analyze the performance of both traditional chord and our proposed three layer hierarchical model we randomly increase the number of nodes in the overlay network and observe the number of hops and time taken to perform the lookup process. Number of superpeers and ultra superpeers are also randomly increased while increasing the number of nodes.

Figure 5 represents the number of hops required to perform the lookup process against increasing number nodes. Blue line shows the number of hops required by traditional chord, whereas green line represents the number of hops used by three layered hierarchical chord to perform the lookup process. Figure shows that as we increase the number of nodes, the number of hops required to perform the query process also increases. However the rate of increase is abrupt in traditional chord as compared to our proposed model. This is because of the fact that traditional chord uses $\frac{1}{2}(\log N)$ hops for lookup, and as the number of nodes increases, the number of hops required to perform a lookup process also increases. Our hierarchical model uses $\frac{1}{2}(\log U + 2)$ hops which significantly

reduces the number of hops required to perform lookups, as the only factor that can increase the number of hops is sufficiently small to produce any significant effect.

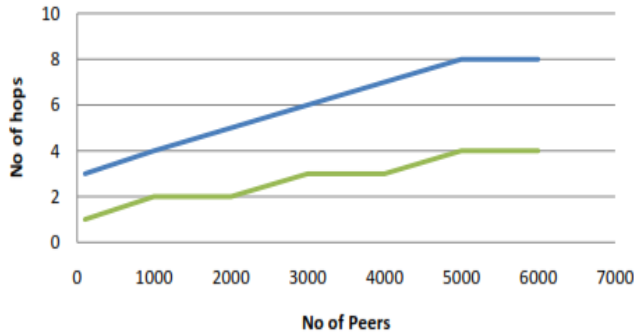


Fig 5: Number of Hops per lookup

Figure 6 shows the lookup latency against increasing number of nodes. It is obvious that as the number of nodes increases the time taken by traditional to perform lookups is greater than that of our hierarchical model. This is because the number of nodes required to perform lookup is less as shown in figure 4, as well as the time taken by each individual node in our hierarchical model is comparatively less than nodes in the traditional chord. This is why, when the number nodes increases, the time taken to perform lookups in traditional chord increases rapidly as compared to our model.

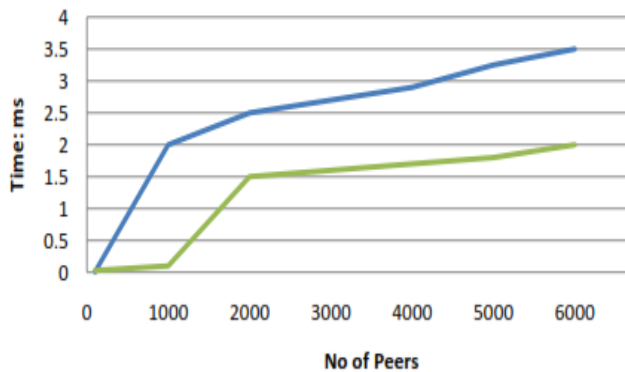


Fig 6: Lookup Latency

VI. CONCLUSIONS

This paper presents a three layer hierarchical overlay architecture, which is efficient, stable, and scalable as shown. Peers in such architecture can frequently join or leave the network, whereas migrating to different layers are also presented in our proposed model. This model is able to discriminate among the peers according to their capability and thus classifying them into three categories: Ultra-Superpeer, Superpeer and Ordinary Peer.

Our hierarchical design offers higher stability by using ultra-superpeers at the upper layer which are more reliable peers. We also presented an instantiation of our 3-layer hierarchical model using Chord at the upper layer and superpeer based single connection intra group structure in middle layer. Since the number of Ultra-Superpeer is less than the number of Superpeer and the number of Superpeer is less than the ordinary peer, the overhead to find a desired data has

become less than two layer architecture by keeping metadata at each layer. To reduce the access load of the database, we used decentralized database to distribute the query load. That approach gives a guarantee for the system to say that there will be no single point of failure. Superpeer and DHT chord based lookup process are used in the hierarchy which also helps to reduce the average number of hops to find the desired peer holding the desired data item.

Lookup latency of the model is also decreased by confirming that the average number of required hops is decreased to find the desired peer. This model executes the lookup process with the help of superpeer and chord protocol that extensively reduces the query traffic load which is common characteristic of the flooding based query. Overall structure ensures that database load is also minimized as the load of a particular peer is split among the neighbor peers. Proposed model offers more stability and scalability than existing chord algorithm, however a lot of work and analysis is required to implement and use this model in practice.

REFERENCES

- [1] Shimul Shil, A.K.M. Mahfuzur Rahman, "A Decentralized Key Database for Overlay Identification", Master's Thesis, Blekinge Institute of Technology, Karlskrona, Sweden.
- [2] Sethom K., Affifi H. and Pujolle G., "Palma: A P2P based Architecture for Location Management", 7th IFIP International Conference on Wireless On-demand Network Systems and Services, Paris, France, 2005.
- [3] Zuji Mao, Douligeris C., "A distributed database architecture for global roaming in next-generation mobile networks", IEEE Journal of Networking, vol. 12, no. 1, pp. 146-160, 2004.
- [4] I. Stoica, R. Morris, and D. Karger, "Chord: A scalable peer-to-peer lookup service for internet application," ACM SIGCOMM Computer Communication, vol. 31, pp. 149-160, 2001.
- [5] Prasanna Ganesan, Krishna Gummadi, and Garchia-Molina H., "Canon in G major: designing DHTs with hierarchical structure", 24th IEEE International Conference on Distributed Computing Systems, Stanford University, CA, USA, February 2005, pp. 263-272.
- [6] Hofstatter Q., Zols S., Michel M., Despotovic Z., and Kellerer W., "Chordella - A Hierarchical Peer-to-Peer Overlay Implementation for Heterogeneous, Mobile Environments", 8th International Conference on Peer-to-Peer Computing, Inst. of Commun. Networks, Tech. Univ. Munchen, Munich, September 2008, pp. 75-76.
- [7] Zoels S., Despotovic Z., Kellerer, W., "Cost-Based Analysis of Hierarchical DHT Design", 6th IEEE International Conference on Peer-to-Peer Computing, Inst. of Commun. Networks, Munich Univ. of Technol., September 2006, pp. 233-239.
- [8] Z. Xu, Y. Hu, "SBARC: A Superpeer Based Peer-to-Peer File Sharing System", Computers and Communications, Eight IEEE International Symposium on Digital Object Identifier, 2003.
- [9] T. Ruixiong, X. Yongqiang, Z. Qian, L. Bo, Y.B. Zhao and L. Xing, "Hybrid Overlay Structure Based on Random Walks", 4th International workshop on Peer-to-Peer Systems, pp. 152-162, 2005.
- [10] Y. Min and Y. Yuanyuan "An efficient hybrid Peer-to-Peer system for distributed data sharing", IEEE International Conference on Parallel and Distributed Processing, pp. 1-10, 2008.
- [11] M. Pandey, S. M. Ahmed and B. D. Chaudhary, "2T-DHT: A Two Tier DHT for Implementing Publish/Subscribe", IEEE International Conference on Computational Science and Engineering, pp. 158165, 2009.
- [12] Rout R.R., Shiva Rama Krishna K. and Kant K., "A Centralized Server Based Cluster Integrated Protocol in Hybrid P2P Systems", 1st IEEE International Conference on Emerging Trends in Engineering and Technology, pp. 390-395, 2008.
- [13] K. Watanabe, N. Hayashibara and M. Takizawa, "A Superpeer-Based Two-Layer P2P Overlay Network with the CBF Strategy", 1st IEEE International Conference on Complex, Intelligent and Software Intensive Systems, pp. 111-118, 2007.

- [14] R. Farha, K. Khavari, N. Abji and A. Leon-Garcia, "Peer-to-Peer Mobility Management for all-IP Networks", IEEE International Conference on Communications, pp. 1946-11952, June 2006.
- [15] P. Zhuo, D. Zhenhua, Q. Jian-jun, C. Yang and L. Ertao, "HP2P: A Hybrid Hierarchical P2P Network", 1st IEEE International Conference on digital society, pp. 18-18, 2007

Automatic Facial Expression Recognition Based on Hybrid Approach

Ali K. K. Bermani
College of Engineering,
Babylon University,
Babil, Iraq

Atef Z. Ghalwash
Computer Science Department,
Faculty of Computers and
Information, Helwan University,
Cairo, Egypt

Aliaa A. A. Youssif
Computer Science Department,
Faculty of Computers and
Information, Helwan University,
Cairo, Egypt

Abstract—The topic of automatic recognition of facial expressions deduce a lot of researchers in the late last century and has increased a great interest in the past few years. Several techniques have emerged in order to improve the efficiency of the recognition by addressing problems in face detection and extraction features in recognizing expressions. This paper has proposed automatic system for facial expression recognition which consists of hybrid approach in feature extraction phase which represent a combination between holistic and analytic approaches by extract 307 facial expression features (19 features by geometric, 288 feature by appearance). Expressions recognition is performed by using radial basis function (RBF) based on artificial neural network to recognize the six basic emotions (anger, fear, disgust, happiness, surprise, sadness) in addition to the natural. The system achieved recognition rate 97.08% when applying on person-dependent database and 93.98% when applying on person-independent.

Keywords- Facial expression recognition; geometric and appearance modeling; Expression recognition; Gabor wavelet; principal component analysis (PCA); neural network.

I. INTRODUCTION

Over the last years, face recognition and automatic analysis of facial expressions has one of the most challenging research areas in the field of computer vision and has received a special importance. The focus of the relatively recently initiated research area of facial expression has lied on sensing, detecting and interpreting human affective states and devising appropriate means for handling this affective information in order to enhance current human machine interface designs. The most expressive way humans display their emotional state is through facial expressions. A human can detect face and recognition on facial expressions without effort, but for a machine and in computer vision it is very difficult. Mehrabian [1] points out that 7% of human communication information is communicated by linguistic language (verbal part), 38% by paralanguage (vocal part) and 55% by facial expression. Therefore, facial expressions are the most important information for emotional perception in face to face communication. Automatic facial expression recognition is an interesting and challenging problem. Automatic recognition of facial expression has wide range of applications in areas such as human computer interaction (HCI), emotion analysis, Psychological area, virtual reality, video-conferencing, indexing and retrieval of image and video database, image understanding and synthetic face animation. In this paper and

most systems of recognition for facial expressions can be divided into three phases. (1) Face detection (2) Feature extraction (facial features). (3) Classification. Detect the face from the image is first important phase. Next and the most important phase is the mechanism for extracting the facial features from the facial image. The extracted facial features are either geometric features such as the shapes of the facial components (eyes, mouth, etc.) and the locations of facial characteristic points (corners of the eyes, mouth, etc.), or appearance features representing the texture of the facial skin in specific facial areas including wrinkles, bulges, and furrows. Appearance-based features include learned image filters from Principal Component Analysis (PCA), Gabor filters, features based on edge-orientation histograms, etc. or hybrid-based include both geometric and appearance features. The final phase is the classifier, which will classify the image into the set of defined expressions.

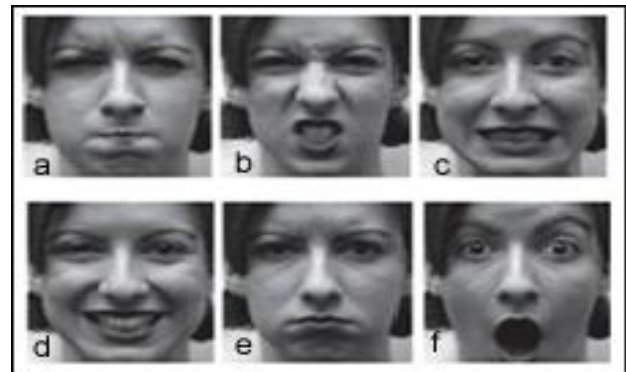


Figure 1. Basic emotion: (a) anger, (b) fear, (c) disgust, (d) joy, (e) sadness, (f) surprise.

There are six universally recognized expressions Angry, Disgust, Fear, Happy, Sad and Surprised (a facial expression into one of the basic facial expressions) as shown in Figure 1. The remainder of this paper is organized as follows: Section II is Related Work on facial expression recognition. Section III, describe Proposed System, In Section IV, experimental results are dataset used in the Proposed System and results. Finally, conclusions are given in Section V.

II. RELATED WORK

Study of facial expressions dates back to 1649, when John Bulwer has written a detailed note on the various expressions

and movement of head muscles in his book “Pathomyotomia“. In 19th century, one of the most important works on facial expression analysis that has a direct relationship to the modern day science of automatic facial expression recognition was the work done by Charles Darwin. In 1872, Darwin [2] wrote a treatise that established the general principles of expression and the means of expressions in both humans and animals. Since the mid of 1970s, different approaches are proposed for facial expression analysis from either static facial images or image sequences. In 1978, Ekman [3] defined a new scheme for describing facial movements. This was called Facial Action Coding Scheme (FACS). FACS combines 64 basic Action Units (AU) and a combination of AU's represents movement of facial muscles and tells information about face expressions. Before the mid of 1990s, facial motion analysis has been used by researchers to perform automatic facial data extraction. Generally existing approaches for facial expression recognition systems can be divided into three categories, based on how features are extracted from an image for classification. The categories are geometric-based, appearance-based, and hybrid-based.

A. Geometric-based approaches

The geometric feature measures the variations in shape, location, distance of facial components like mouth, eyes, eyebrows, nose, etc. In their earlier work Kobayashi and Hara [4], [5], proposed a geometric face model of 30 facial characteristic points FCPs. In their later work Kobayashi and Hara [6], use a CCD camera in monochrome mode to obtain a set of brightness distributions of 13 vertical lines crossing the FCPs. First, they normalize an input image by using an affine transformation so that the distance between irises becomes 20 pixels. From the distance between the irises, the length of the vertical lines is empirically determined. The range of the acquired brightness distributions is normalized to [0,1] and these data are given further to a trained neural network NN for expression emotional classification. Pantic et al.[7] proposed automated method for detecting facial actions by analyzing the contours of facial components, including the eyes and the mouth. A multi-detector technique is used to spatially sample the contours and detect all facial features. Arule-based classifier is then used to recognize the individual facial muscle action units(AUs). They achieve 86% recognition rate.

B. Appearance-based approaches

The appearance features present the appearance (skintexture) variations of the face, such as wrinkles, furrows, etc. It can be extracted on either the whole face image or specific regions in a facial image. Feng [8] uses Local Binary Patterns (LBP) to extract facial texture features and combines different local histograms to recover the shape of the face. Classification is performed through a coarse-to fine scheme, where seven templates were designed to represent the corresponding seven basic facial expressions. At first, two expression classes are selected based on the distance from the test image to the seven templates. The final classification is then done via a K-nearest neighbor classifier with the weighted Chi-square statistic. Deng et al.[9] proposed a facial expression recognition system based on the local Gabor filter bank. They selected some parts of the frequency and orientation parameters instead of using the entire global filter

bank to extract the features, and tested them on the JAFFE facial database which stores posed facial expressions. In order to select a few Gabor filters, they covered all the frequencies and orientations, but only selected one frequency for each orientation or each increase in the interval between neighboring frequencies with the same orientation. Lajevardi et al [10] they extract features from a picture of the whole face. The features are generated using the log-Gabor filters. The log-Gabor filters can be constructed with an arbitrary bandwidth that can be optimized to produce a filter with minimal spatial extent. Five scales and eight orientations are implemented to extract features from face images. This leads to 40 filter transfer functions $\{H_1, H_2, \dots, H_{40}\}$ representing different scales and orientations. The objective is to choose only the log-Gabor filters which minimize the absolute value of the spectral difference between the original image filter output and the noisy image filter output. Initially, the error is calculated for all of the training images, and for each image a set of 4 log-Gabor filters with the smallest value of the spectral difference is selected. The selected optimal subset of 4 log-Gabor filters allowed reducing the features dimension from 40 to only 4 arrays of size 60×60 for each image. The feature arrays are transferred to feature vectors of length 3600 for each image. Murthy and Jadon [11] propose a modified PCA (eigenspaces) for Eigen face reconstruction method for expression recognition. They have divided the training set of Cohn kanade and JAFFE databases into 6 different partitions and Eigen space is constructed for each class, then image is reconstructed. Mean square error is used as a similarity measure for comparing original and reconstructed image.

C. Hybrid-based approaches

Hybrid features uses both geometric and appearance based approaches to extract features. Zhang et al. [12] presented a Facial Expression Recognition system where they have compared the use of two types of features extracted from face images for recognizing facial expression. Geometric positions of set of fiducial points (34 facial points) and multiscale & multi orientation gabor wavelet coefficient extracted from the face image at the fiducial points are the two approaches used for feature extraction. These are given to neural network classifier separately or jointly and results are compared. Their system achieves a generalized recognition rate of 73.3% with geometric positions alone, 92.2% with Gabor wavelet coefficients alone, and 92.3% with the combined information.

Zhang and Ji [13] proposed a multi-feature technique that is based on the detection of facial points, nasolabial folds, and edges in the forehead area. In their method, facial features are extracted by associating each AU with a set of movements, and then classified using a Bayesian network model. Jun Ou et al.[14] presents the Automatic Facial Expression Recognition (AFER) using 28 facial feature points and Gabor wavelet Gabor wavelet filter provided with 5 frequencies, 8 orientations for facial feature localization, PCA for feature extraction and K nearest neighbor (KNN) for expression classification. Their system achieves a generalized recognition rate of 80%. Monwar & Rezaei [15] use location and shape features to represent the pain information. These features are used as inputs to the standard back-propagation in the form of a three-layer neural network with one hidden layer for

classification of painful and painless faces. They achieve 91.67% recognition rate using 10 hidden layer units.

III. PROPOSED SYSTEM

This paper proposes an automatic system for facial expression recognition, a system will address the problems of facial expression recognition about how to detect a human face in static images and how to represent and recognize facial expressions presented in those faces by using a hybrid approach for facial feature information extraction and then Classification of the facial expressions into the six basic emotions (anger, disgust, fear, happy, sad and surprise) in addition to the neutral one. The system operates after entering the image that contains the face which will begin The first phase; face detection process and segment the face image, Then extract feature vector (geometric feature and appearance feature) by apply the feature extraction process on face image, The final phase; uses this feature as an input into the radial basis function artificial neural network to recognize the facial expressions. Figure 2, shown the block diagram of the proposed system.

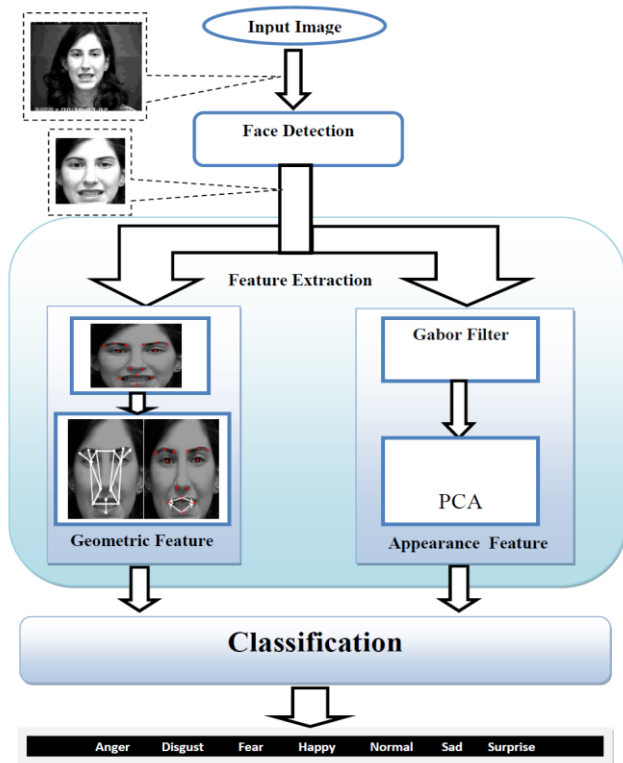


Figure 2. Block diagram of the facial expression recognition system.

A. Face Detection

In most systems of facial expression recognition the first phase is the face detection and segmentation from the image. The purpose of face detection is to determine the presence and the position of face in an image. Many algorithms have been proposed for face detection, e.g. the Viola-Jones face detector [16]. The open source code library (OpenCV) that employs algorithm based on Viola & Jones features Sreekar, Krishna [17],[16], used in this system. Sreekar [17] Works to determination face image, this system will segment the face

image the specific. 300x300 pixels is uniform size of face image that resulted from the Open CV face detector. Figure 3, shows example of the Open CV face detector.



Figure 3. An example of the OpenCV face detector.

B. Facial Features Extraction

The second phase and the most important in the systems of facial expression recognition is the facial feature extraction which is based on finding features that conveying the facial expression information. These problems can be viewed as a dimensionality reduction problem (transforming the input data into a reduced representation set of features which encode the relevant information from the input data). This system uses two approaches to extract the facial features: geometric and appearance features.

1) *Geometric Features Extraction*: In Geometric approach, 19 features are extracted from the face image. In this phase the same approach that Youssif et al. [18] will be used. They extracted 19 features from the face image by executing the segmentation process to divide the face image into three regions: mouth, nose and two eyes and two eyebrows and then located the facial characteristic points (FCPs) in each face component by using mouth, nose, eyes and eyebrows FCPs extraction techniques. Finally; certain lengths between FCPs are calculated using Euclidean distance D:

$$D = \sqrt{(x1 - x2)^2 + (y1 - y2)^2} \quad (1)$$

Where (x1, y1) and (x2,y2) are the coordinates of any two FCPs P1(x1,y1) and P2(x2,y2) respectively. Also two angles in the mouth area are computed to represent with geometric lengths the geometric features, typically results are in Figure 4 (a, b).

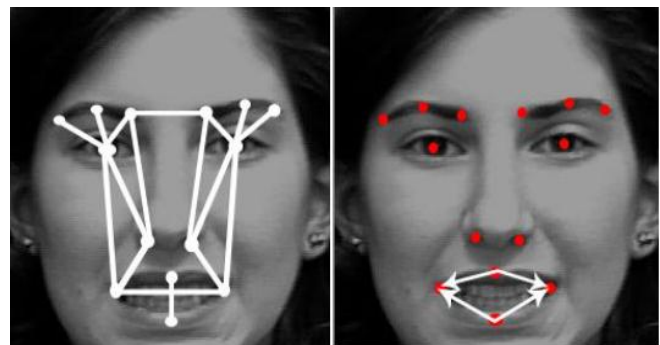


Figure 4. Geometric features: (a) geometric lengths, (b) mouth angles

2) *Appearance Features Extraction*: The second approach used to extract feature is appearance approach. Various feature extraction methods can be used to extract facial expression features such as, the active appearance models (AAM), eigenfaces, and Gabor wavelets that proposed in this paper. The Gabor wavelets reveal much about facial expressions as both transient and intransient facial features that often give rise to a change in contrast.

a) *pre-processing and Gabor filters*: After face detection process, 300x300 pixels is the size of face image, will be done by using cropping process on face image to be 200x300 pixels and then do resize to the cropped image to be 200x200. After pre-processing process of face image will use Gabor filter to extract features. Gabor wavelet represents the properties of spatial localization, orientation selectivity, and spatial frequency selectivity. The Gabor wavelet can be defined as

$$\Psi(\mathcal{Z}) = \frac{p_{u,r}^2}{\sigma^2} \exp\left(-\frac{p_{u,r}^2 \mathcal{Z}^2}{2\sigma^2}\right) \left[\exp(ip_{u,r}\mathcal{Z}) - \exp\left(-\frac{\sigma^2}{2}\right) \right] \quad (2)$$

Where $\mathcal{Z} = (x,y)$, and u and r define the orientations and scales of the Gabor wavelet, respectively [19]. $p_{u,r}$ is defined as

$$P_{u,r} = P_r e^{i\Phi_u} \quad (3)$$

Where $P_r = p_{\max}/f^r$ and $\Phi_u = \pi u/6$. P_{\max} is the maximum frequency, and f is the spacing factor between kernels in the frequency domain. In this experiment the Gabor wavelet with three scales ($r = 0,1,2$) and six orientations ($u = 0,1,2,3,4,5$), and parameters $\sigma = 2\pi$, $p_{\max} = \pi/2$, and $f = \sqrt{2}$ are applied. This reproduces 18 different Gabor kernels (3 scales * 6 orientations). The Gabor representation of an image, which is called a Gabor image, is produced from the convolution of the image with the Gabor kernels as defined by Equation 2. For each image pixel, it produces 18 Gabor images and each one consists of two Gabor parts: the real part and the imaginary part. Thereafter, these two parts are transformed into two kinds of Gabor feature: the magnitude and the phase. In this study only the magnitude features are used to represent the facial Gabor features. Changing the facial positions will cause only slight variations in the magnitude of the Gabor wavelet, while the phase of the Gabor wavelet output is very sensitive to the position of the face [20]. The 200x200 pixels in the Gabor image produce a large number of features is 40000 feature vectors. To produce a smaller number of features, the image size has been re-sized to 25x25 pixels, resulting in only 625 features.

b) *Dimension reduction with Principal component analysis (PCA)*: After re-sizing the Gabor image to 25x25 pixels the image is further compressed, resulting in 625x1 feature vectors for each image. PCA is a way of identifying patterns in the data and expressing the data in such a way as to highlight their similarities and differences [21]. Since the patterns in the data can be difficult to identify in data of high dimensions where the graphical representation is not available, PCA is a powerful tool for analyzing such data. The other

main advantage of PCA is its ability to reduce the number of dimensions without much loss of information. PCA will be used in this experiment to reduce the features of each image (its 18 Gabor images). To perform PCA, we reshape each Gabor image into one column producing a matrix have 625 rows and 18 columns, then the mean value of these Gabor images columns are subtracted from each Gabor image column. The subtracted means are the average across each dimension, and then the covariance matrix is calculated (which is 18*18 matrix). Then, based on the covariance matrix, eigenvectors and eigenvalues are calculated. Next, the 16 most significant eigenvalues are selected to obtain the eigenvectors (then will have a matrix of 16 columns and 18 rows). These eigenvectors are input into the radial basis function neural network classifier after reshaping it into 1 row and adding the geometry features to it (16*18=288 appearance features + 19 geometry features).

C. Facial Expression Classification

After extract feature from face image, the next and last phase in this system is the classification of the facial expressions. The facial feature vector of 307 feature elements (19 geometric features plus 288 appearance features) forms the input of radial basis function RBF neural network. As shown in Figure 5. The RBF neural network has one hidden layer uses neurons with RBF activation functions (ϕ_1, \dots, ϕ_m) for describing the receptors. The following equation is used to combine linearly the outputs of the hidden neurons:

$$y = \sum_{j=1}^m w_j \phi_j(\|x - t_j\|) \quad (4)$$

Where $\|x - t\|$ is the distance of the input vector $x = (x_1, \dots, x_{m1})$ from the vector $t = (t_1, \dots, t_{m1})$ which is called the center for m inputs (receptors). The proposed system use the Gaussian function described by the equation 5:

$$\phi(x) = e^{-\frac{x^2}{2\sigma^2}} \quad (5)$$

Where σ is the spread parameter of the Gaussian function. Number of neurons is 250 and RBF spread value is 250 used in this system with person-dependent datasets and Number of neurons is 80 and RBF spread value is 300 used in this system with person-independent datasets.

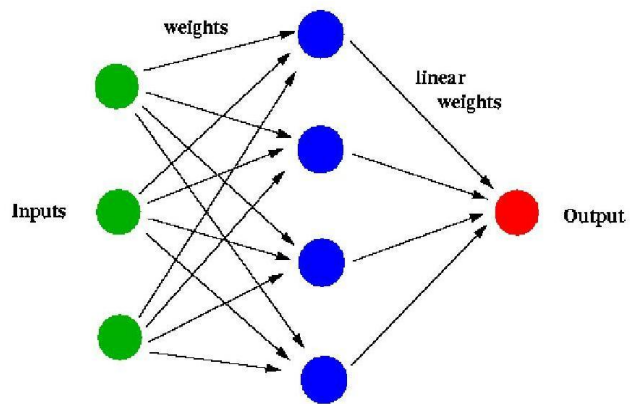


Figure 5. Radial basis function neural network architecture

IV. EXPERIMENTS AND RESULTS

The performance of proposed system is analysed on the Cohn-Kanade facial expression database Anitha et al. [22], which is commonly used in research on facial expression recognition. This database includes approximately 2000 image sequences with 640x480 pixels arrays and 8-bit precision for grayscale values from over 200 subjects. Subjects in the available portion of the database are 97 university students. They are ranged in age from 18 to 30 years. 65% are females, 15% are African-American, and 3% are Asian or Latino, Figure 6, Samples from original Cohn-Kanade Facial Expression Database. In order to compare the results of this system with that of Youssif et al [18], the database used in [18], has been adopted. They prepared two datasets from the available portion to examine the proposed system with different two manners. In the first manner the proposed system trained on each subject (person) different facial expression. So; all subjects with their facial expressions were exist in both training and testing datasets which called person-dependent dataset. In this dataset the last five images from each subject facial expression were taken and used the odds (60%) in training and the evens (40%) in testing. In the second manner the system trained in the facial expressions independent on the subjects. The person-independent dataset were prepared by selecting one image that convey facial expression information from each subject facial expressions images to form the seven facial expression classes. Then a 60% of the images were selected for training phase and the rest (40%) for testing. In both datasets the normal class was prepared by picking up the first image from each expression for subject.



Figure.6: Samples from original Cohn-Kanade Facial Expression Database.

The results in tables (1,2) show that the proposed system can classify the facial expressions correctly with recognition rates 97.08% using person-dependent dataset and recognition rate 93.98% using person-independent dataset. From these results we can conclude that the proposed system has improved recognition rate in both database used with the system used by Youssif et al [18], where the percentage was 96 using person-dependent dataset and 93 using person-independent dataset. The system is implemented in MATLAB 7.7.0 on 1.73GHz Microsoft Windows XP professional workstation.

%	Anger	Disgust	Fear	Happy	Normal	Sad	Surprise
Anger	100	0	0	0	0	0	0
Disgust	0	92.5	2.5	0	2.5	2.5	0
Fear	0	0	100	0	0	0	0
Happy	0	2.63	5.26	92.1	0	0	0
Normal	0	0	0	0	100	0	0
Sad	0	2.5	0	0	2.5	95	0
Surprise	0	0	0	0	0	0	100
Recognition rate	97.08						

%	Anger	Disgust	Fear	Happy	Normal	Sad	Surprise
Anger	93.33	0	0	0	6.66	0	0
Disgust	7.14	84.6	7.14	0	0	0	0
Fear	0	0	93.33	0	6.66	0	0
Happy	0	0	0	100	0	0	0
Normal	0	0	0	0	100	0	0
Sad	0	0	6.66	0	6.66	86.66	0
Surprise	0	0	0	0	0	0	100
Recognition rate	93.98						

V. CONCLUSIONS

This paper proposed a method for automatic facial expression recognition. A system is capable detect a human face in static images and extract features by a hybrid approach (geometric and appearance approach) and then classify expressions presented in those faces. A hybrid approach is used for facial features extraction as a combination between geometric an appearance facial features. Radial Basis Function RBF based neural network is used for facial expression recognition.

The proposed system can classify the facial expressions correctly with classification rates between 92% and 100% (recognition rate 97.08%) using person-dependent dataset and between 84.6% and 100% (recognition rate 93.98%) using person-independent dataset. The proposed system could classify anger, fear, normal, sad and surprise in maximum rates but the disgust and happy in minimum rate in person-dependent dataset. In person-independent dataset the anger, fear, happy, normal and surprise classes could be recognized at maximum rates but fear and sad classes registered the minimum recognition rate.

REFERENCES

- [1] A. Mehrabian, "Communication without Words," *Psychology Today*, Vo.1.2, No.4, pp 53-56, 1968.
- [2] C. Darwin, "The expression of the emotions in man and animal," J.Murray, London, 1872.
- [3] P. Ekman and W.V. Friesen, "Facial Action Coding System (FACS)," Manual. Palo Alto: Consulting Psychologists Press, 1978.
- [4] H. Kobayashi and F. Hara, "Recognition of Six Basic Facial Expressions and Their Strength by Neural Network," *Proc. Int'l Workshop Robot and Human Comm*, pp. 381-386, 1992.

- [5] H. Kobayashi and F. Hara, "Recognition of Mixed Facial Expressions by Neural Network," *Proc. Int'l Workshop Robot and Human Comm.*, pp. 387-391, 1992.
- [6] H. Kobayashi and F. Hara, "Facial Interaction between Animated 3D Face Robot and Human Beings," *Proc. Int'l Conf. Systems, an, Cybernetics*, pp. 3,732-3,737, 1997.
- [7] M. Pantic and L. J. M. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(3):1449-1461, 2004.
- [8] X. Feng, "Facial expression recognition based on local binary patterns and coarse-to-fine classification," *In Proceedings of Fourth International Conference on Computer and Information Technology*, pages 178-183, 2004.
- [9] HB. Deng, LW. Jin, LX. Zhen and JC. Huang, "A new facial expression recognition method based on local Gabor filter bank and PCA plus LDA," *Int J Inf Technol 11(11)*:86-96. 2005.
- [10] S. Lajevardi and M. Lech, "Facial expression recognition from image sequences using optimized feature selection," *Image and Vision Computing New Zealand, IVCNZ 2008. 23rd International Conference*, pp. 1-6, 2008.
- [11] G. R. S. Murthy, R. S. Jadon, "Effectiveness of Eigenspaces for facial expression recognition," *International Journal of Computer Theory and Engineering*, Vol. 1, No. 5, 1793-8201, pp. 638-642. December 2009.
- [12] Z.Zhang, M. Lyons, M. Schuster and S. Akamatsu "Comparison between geometry-based and Gabor wavelets-based facial expression recognition using multi-layer perceptron," *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, pp. 454-459.1998.
- [13] Y. Zhang and Q. Ji. "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):699-714, 2005.
- [14] Jun Ou, Xiao-Bo Bai, Yun Pei, Liang Ma, Wei Liu, "Automatic facial expression recognition using Gabor filter and expression analysis," *Second International Conference on Computer Modeling and Simulation*, IEEE, pp 215-218.2010.
- [15] M. Monwar, and S. Rezaei, "Pain recognition using artificial neural network," *Signal Processing and Information Technology, 2006 IEEE International Symposium on*. 2006.
- [16] P. Viola, and M. Jones. "Robust real-time object detection," *Second international workshop on statistical and computational theories of vision-modeling, learning, computing, and sampling*. 2001.
- [17] Sreekar Krishna. Open CV Viola-Jones Face Detection in Matlab. [Online] available :<http://www.mathworks.com/matlabcentral/fileexchange/19912> .June 6, 2010.
- [18] A. Youssif, and W. Asker, "Automatic Facial Expression Recognition System Based on Geometric and Appearance Feature," *Computer and Information Science*, Vol.4, No.2, March 2011.
- [19] TS. Lee "Image representation using 2D Gabor wavelets," *IEEE Trans Pattern Anal Mach Intell* 18:959-971, 1996.
- [20] I. Kotsia, I. Bucu and I. Pitas. "An analysis of facial expression recognition under partial facial image occlusion," *J Image Vision Computer*, 26:1052-1067, 2008.
- [21] LI. Smith, "A tutorial on principal component analysis," *Cornell University*, pp 2-22, 2002.
- [22] M. Anitha, K. Venkatesha and B. Adiga, "A survey on facial expression databases," *International Journal of Engineering Science and Technology*, Vol. 2(10), 5158-5174, 2010.

Design of A high performance low-power consumption discrete time Second order Sigma-Delta modulator used for Analog to Digital Converter

Radwene LAAJIMI

University of Sfax
Electronics, Micro-technology and Communication
(EMC) research group
Sfax (ENIS), BP W, 3038 Sfax, Tunisia

Mohamed MASMOUDI

University of Sfax
Electronics, Micro-technology and Communication
(EMC) research group
Sfax (ENIS), BP W, 3038 Sfax, Tunisia

Abstract—This paper presents the design and simulations results of a switched-capacitor discrete time Second order Sigma-Delta modulator used for a resolution of 14 bits Sigma-Delta analog to digital converter. The use of operational amplifier is necessary for low power consumption, it is designed to provide large bandwidth and moderate DC gain. With 0.35 μ m CMOS technology, the $\Sigma\Delta$ modulator achieves 86 dB dynamic range, and 85 dB signal to noise ratio (SNR) over an 80 KHz signal bandwidth with an oversampling ratio (OSR) of 88, while dissipating 9.8mW at ± 1.5 V supply voltage.

Keywords- CMOS technology; Analog-to-Digital conversion; Low power electronics; Sigma-Delta modulation; switched-capacitor circuits; transconductance operational amplifier.

I. INTRODUCTION

Analog to digital converters (ADC) with high resolution are widely used in the areas of instrumentation, measurement, telecommunications, digital signal processing, consumer electronics, etc. With the advancements in the Very Large Scale Integration (VLSI) technologies, the focus is shifted on oversampling and $\Delta\Sigma$ converters for applications requiring high precision analog-to digital conversion with narrow bandwidth [1, 2, 3, 4]. They are preferred because of their inherent relaxed sensitivity to analog circuit errors and reduced analog processing as compared to other analog-to-digital conversion techniques. These advantages come at the expense of relatively large amount of digital processing and the working of the majorpart of circuit at a clock rate which is much higher than the analog-to-digital conversion rate. Because of using higher clock rates and feedback loop, these converters tend to be robust in the face of analog circuit imperfections [3] and do not require trimmed components which are considered necessary in conventional high precision Nyquist rate ADCs.

For these reasons, high precision $\Delta\Sigma$ ADCs can be implemented using high density VLSI processes.

$\Sigma\Delta$ ADC is a system which consists of a $\Sigma\Delta$ modulator followed by a digital decimation filter. The $\Sigma\Delta$ modulator over-samples the input signal i-e performs sampling at a rate much higher than the nyquist rate. The ratio of this rate to the Nyquist rate is called over-sampling ratio (OSR).

After over-sampling, it typically performs very coarse analog-to-digital conversion at the resulting narrow-band signal. By using coarse digital-to-analog conversion and feedback, the quantization error introduced by the coarse quantizer is spectrally shaped i-e the major portion of the noise power is shifted outside the signal band. This process is called quantization noise shaping. The digital decimation filter removes the out-of-band portion of the quantization error and brings back the output rate to Nyquist rate.

The paper is organized as follows. Section II presents a review of $\Sigma\Delta$ modulator with theoretical analysis. Section III presents the described structure of the second order $\Delta\Sigma$ modulator with simulations results. In section IV, all main parameters of the proposed modulator are indicated with a full comparison of the most popular designs in which the performance of each modulator is cited in table IV. Conclusion is drawn in Section V.

II. REVIEW OF SIGMA-DELTA MODULATOR

Signal-to-noise ratio (SNR) and Dynamic range (DR) are two most important specifications commonly used to characterize the performance of over-sampling sigma-delta ADCs [5]. The system design begins with the calculation of the dynamic range. According to the design theory, the DR of a sigma-delta modulator can be gotten by following formula.

$$DR(dB) = 10 \log_{10} \left[\frac{3}{2} \frac{(2L+1)}{\Pi^{2L}} OSR^{2L+1} (2^B - 1)^2 \right] \quad (1)$$

The theoretical DR is a function of the modulator order L, oversampling ratio OSR, and the numbers of bits in the quantizer B.

The expression (1) reveals that an additional bit in the internal quantizer can roughly obtain a 6-dB improvement of DR. This improvement is independent of the OSR, while the improvement obtained with increasing the order L is dependent on it. The DR of a theoretical L th-order Sigma-Delta converter increases with OSR in (L + 1/2) bits/octave. This is shown in Figure 1, where the DR is plotted as a function of the oversampling ratio and the modulator order, in case of a single-bit internal quantizer.

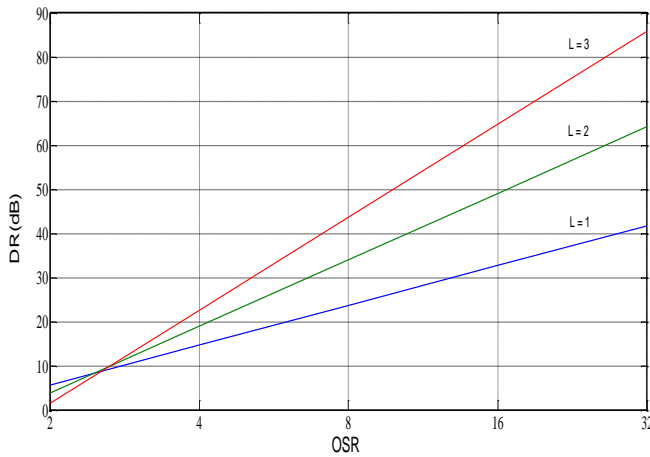


Figure 1. DR Vs OSR of Lth-order theoretical $\Sigma\Delta$ Modulators

The SNR of a converter is the ratio of the input signal power to the noise power measured at the output of the converter. The maximum SNR that a converter can achieve is called peak signal-to-noise-ratio (SNR_p). The noise here should include the quantization and circuit noise. The SNRP of the L-th order Sigma-Delta modulator can be calculated as:

$$SNR_p = \frac{3\Pi}{2} (2^B - 1)^2 (2L + 1) \left(\frac{OSR}{\Pi}\right)^{2L+1} \quad (2)$$

The SNR of the $\Sigma\Delta$ ADC can be increased by $(2L + 1) 3dB$, or $L + 0,5bits$ by doubling the oversampling ratio, where L denotes the order of the loop filter. It is tempting to raise the oversampling ratio to increase the SNR of the Sigma-Delta modulator. However, it is restricted by the speed limit of the circuit and the power consumption.

In practice, for the same performance, it is preferred to lower the oversampling ratio. Another driving force is the ever increasing bandwidth requirement, which also needs to lower the oversampling ratio.

For high bandwidth converters, the oversampling ratio should be kept as low as possible. A lot of efforts have been made at the system level to lower the oversampling ratio and maintain the same performance.

Starting from the desired 14-bit of resolution, the Sigma-Delta ADC was designed. The DR is computed as follows:

$$DR^2 = 3 (2^{2 \times N - 1}) \quad (3)$$

$$DR^2 = 3 (2^{2 \times 14 - 1}) \quad (4)$$

Where N represents number of bits, hence DR (dB) is equal to 86.04dB. For computing the OSR for first, second and third order devices, a small MATLAB script was developed using the following equation:

$$OSR = \left(\frac{2}{3} \frac{DR^2 \Pi^{2L}}{2L + 1} \right)^{\frac{1}{2L + 1}} \quad (5)$$

OSR means Over Sampling Ratio which is based on the following formula:

$$OSR = \frac{F_s}{2 \times f_b} \quad (6)$$

If we choose an OSR equal to 88, F_s is the sampling frequency which is calculated to 14.08MHz, and f_b means base band frequency (80 KHz). Table I shows the results of the equation (4) for a first, second and third order.

TABLE I. OSR FOR FIRST TO THIRD MODULATOR'S ORDER

Modulator's order	OSR
1 st	960
2 nd	88
3 rd	24

For a first order modulator an OSR of 960 is needed, given a relatively high sampling frequency for the 0.35 μ m CMOS ultra-low-voltage system. For the second and third order devices, better sampling frequencies are required. Although for the third order modulator, the lowest sampling frequency is needed, this frequency implies the use of higher consumption and taking up more area. For that reason, a second order modulator was chosen for this application.

N is the effective number of bits of $\Delta\Sigma$ converter which is given by [6]:

$$N = \frac{1}{2} \log_2 \left[(2B - 1)^2 (2L + 1) OSR^{2L+1} / (\pi)^{2L} \right] \quad (7)$$

Where B represents number of bits of the quantization circuitry ($B = 1$).

In this case, OSR is equal to 88, F_s is the sampling frequency which is calculated to 14.08MHz. We obtained an effective number of bits equal to 14 bits which verify the two last results in equations (1) and (3).

III. PROPOSED SECOND ORDER SIGMA -DELTA MODULATOR

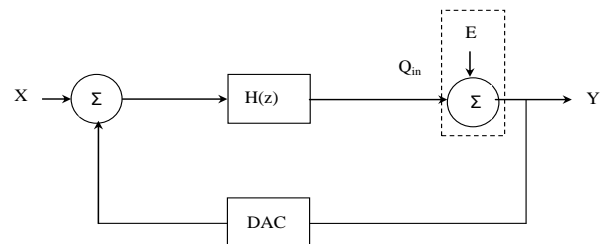


Figure 2. Linear model of conventional first order $\Sigma\Delta$ Modulator

As shown in figure 2, first-order $\Sigma\Delta$ modulator has the advantages of being simple, robust and stable. Despite these good points, its overall performance in terms of resolution and idle-tone generation is inadequate for most applications. Second-order $\Sigma\Delta$ modulator overcomes these disadvantages at the expense of increased circuit complexity and reduced signal range. According to figure 2, the modulator can be considered as a two-input $[E(z)$ and $X(z)]$, one-output $[Y(z)]$ linear system. The output signal is expressed as:

$$Y(z) = STF(z).X(z) + NTF(z).E(z) \quad (8)$$

Where $STF(z)$ is the signal transfer function and $NTF(z)$ is the noise transfer functions, which are given by:

$$STF(z) = \frac{Y(z)}{X(z)} = \frac{H(z)}{1+H(z)} \quad (9)$$

$$NTF(z) = \frac{Y(z)}{E(z)} = \frac{1}{1+H(z)} \quad (10)$$

By using superposition principle, the output signal is obtained as the combination of the input signal and the noise signal, with each being filtered by the corresponding transfer function:

$$Y(z) = STF(z).X(z) + NTF(z).E(z) \quad (11)$$

If we choose $STF(z)$ equal to Z^{-1} and $NTF(z)$ equal to $1-Z^{-1}$ we obtain:

$$STF(z) = z^{-2} \quad (12)$$

$$NTF(z) = (1-z^{-1})^2 \quad (13)$$

In this case, the output signal for the ideal linear model can be written as:

$$Y(z) = X(z).z^{-2} + E(z).(1-z^{-1})^2 \quad (14)$$

To achieve the objective cited above, All main parameters of the described modulator are summed up in Table II.

TABLE II. DESIGNED MODULATOR PARAMETERS

Parameters	Value
Order of modulator	2
Sampling Frequency (clock)	14.08 MHz
Signal Band width	80 KHz
Over sample Ratio(OSR)	88
Resolution	14 bits

The functional diagram of the second order modulator simulated using Simulink in MATLAB is shown in Figure 3. The single bit DAC is replaced by a simple wire. The input is a sinusoidal signal with 0.4 V amplitude and frequency 20 kHz. This signal is fed through two integrators and is connected to the comparator at the output.

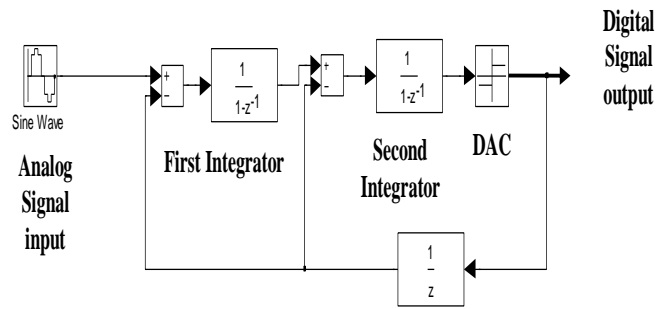


Figure 3. Block diagram of Second Order $\Sigma\Delta$ Modulator

The modulated output as seen through the scope is shown in Figure 4 with the input signal overlaid on it.

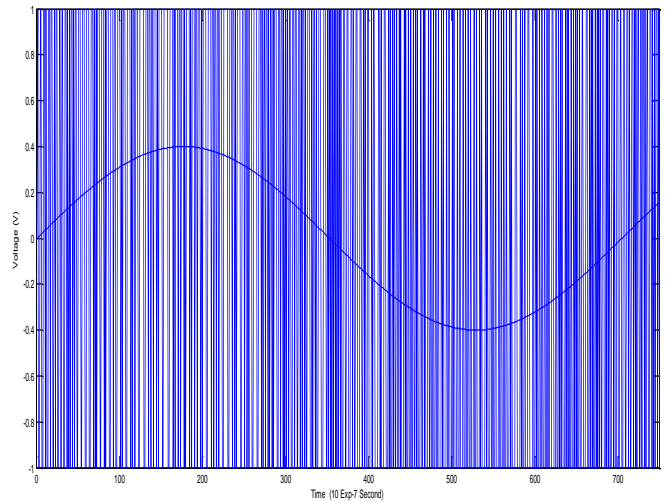


Figure 4. Pulse density output from a sigma-delta modulator for a sine wave input

A discrete Fourier Transform (DFT) of the sampled output signal is performed to calculate the SNR of the system. The logarithm of the amplitude of the signal is plotted versus the signal frequency and the SNR is found to be close to 85.17 dB as shown in Figure 5.

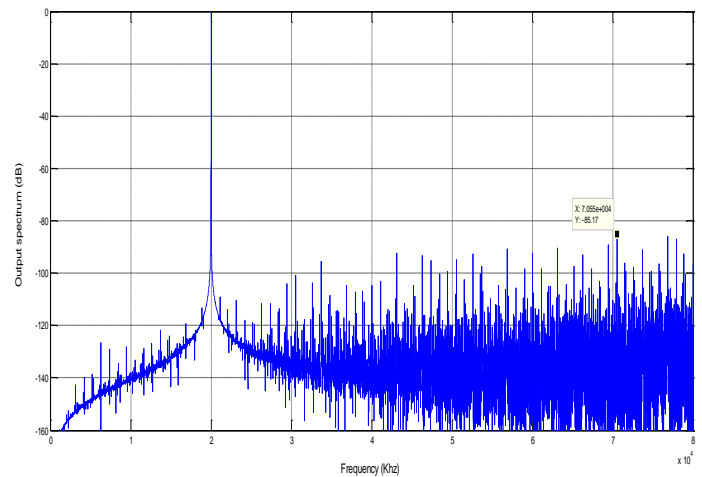


Figure 5. Frequency spectrum zoom of the modulated signal

It can be seen that second order noise shaping is taking place wherein most of the noise is pushed to the higher frequency bands as shown in figure 6. The original signal can be retrieved using a digital low pass filter.

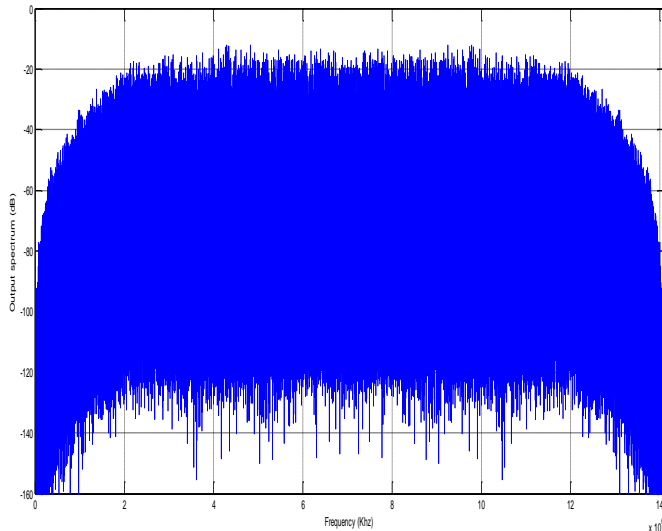


Figure 6. Frequency spectrum of the modulated signal

Figure 7 shows a block diagram of a complete Second-order $\Sigma\Delta$ modulator. It is made up of two integrators, a comparator, and digital to analog converter (DAC). These include switches Q and Q' for applying one of two reference node voltages, $+V_{ref}$ and $-V_{ref}$, depending on comparator output polarity. The two integrators are based on amplifier, and use two phase, with the respective phases denoted by ϕ_1 and ϕ_2 .

As shown in figure 7, signal sampling is completed by the first integrator connecting to the input. For this reason, the amplifier determines the whole performance of the sigma-delta modulator and needs to be carefully designed. Each integrator is made up of one operational amplifier. This operational amplifier has an adequate bandwidth and higher voltage gain. It is composed of two input transistors formed by P-channel MOSFETs M1 and M2 in order to reduce 1/f noise [7].

This stage of op-amp is also formed by N-channel MOSFETs, M3 and M4. The use of transistor M7 as a P-channel common source amplifier forms the second stage of op-amp. The polarisation block is formed by ten transistors (M5, M6, M8, M9, M10, M11, M12, M13, M14 and M15) with variable input voltage V_{in} and input resistance R_{in} .

The overall gain of the amplifier is found to be given by:

$$A_v = g_{m1}(r_{ds2} // r_{ds4}) \cdot g_{m7}(r_{ds6} // r_{ds7}) \quad (15)$$

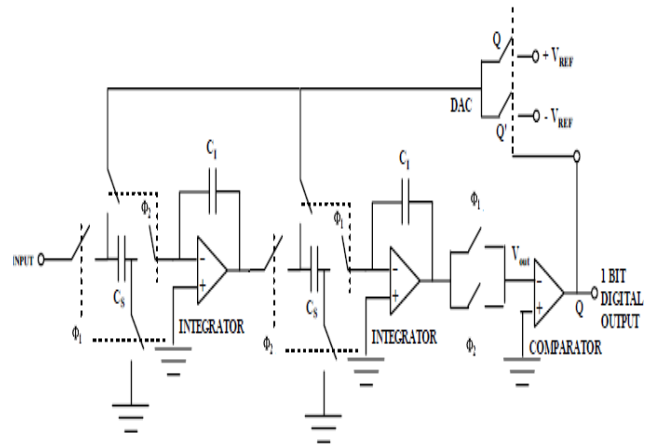


Figure 7. A complete Second-order $\Sigma\Delta$ modulator

Where g_{mi} is the transconductance and r_{dsi} is the drain to source resistance of i^{th} transistors with $i = 1, 2, 4, 6, 7$.

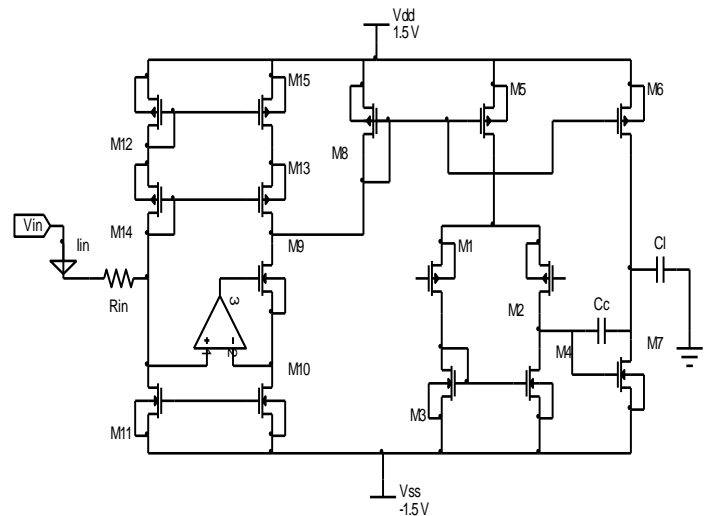


Figure 8. Miller Operational Transconductance Amplifier

Considering the appeal to moderate speed and noise of operational amplifier together with the low power consumption of the operational amplifier, Miller structure is chosen, and the structure of circuit is shown in figure 8. The system has problems of instability because the signal will pass through two-stage circuit and may bring extra pole and zero in this operational amplifier, so it requires importing frequency compensation. Simulations results using T-spice shows that output frequency response of our operational amplifier achieves a gain of 60 dB with a large gain bandwidth of 82 MHz and a phase margin of 62° to ensure more stability (figure 9).

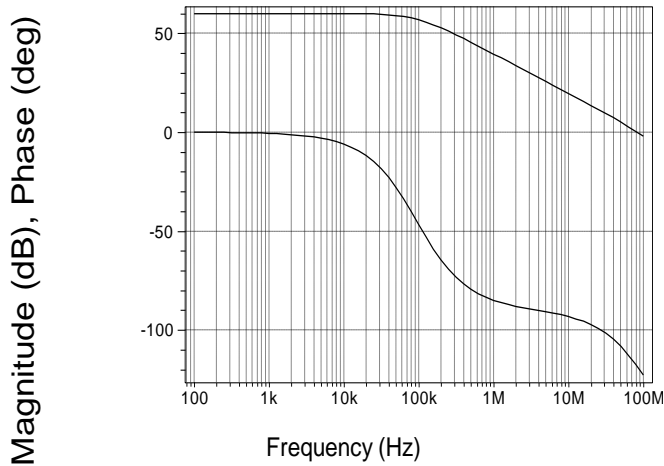


Figure 9. Frequency response of Operational Transconductance Amplifier

A. Why high performance and low power?

With the rapid development of computers and communications, a very large number of chips required to have higher performance, low power and small size. Hence analogue, mixed signals with low power and especially Sigma-Delta modulators become more and more important. Operational amplifier structure for almost all modulators circuits determine the performance of analogue structure, which largely depends on their characteristics. One of the popular op-amp is a two-stage Miller op-amp. It introduces an important concept of compensation. The object of this compensation is to maintain stability of the operational amplifier. This last is made up of three stages even though it is often referred to as a “two-stage” op-amp, ignoring the buffer stage. The first stage is composed of the input devices of the differential pair which are formed by P-channel or N-channel MOSFETs. It plays a very important role due to its differential input to single-ended output conversion and its high gain. In addition this stage of op-amp also had the current mirror circuit formed by N-channel MOSFETs. In the other hand the second stage is formed by only one transistor which serves as a P-channel common source amplifier. The current I_{bias} of the op-amp circuit goes through current mirrors formed by the rest of P-channel transistors in order to produce a low current of $10\mu A$ [17].

According to figure 10, the DC characteristic for the V-I converter for different values of resistance ($R_{out}=100\ \Omega$, $R_{out}=2.5\ K\Omega$, $R_{out}=5\ K\Omega$) is presented. In the one hand the full input voltage swing capability is evident with truly linearity. In the other hand, in order to ensure low and low consumption an input voltage V_{in} varied from -1.1V to 0V to provide a current from $-10\mu A$ to $150\mu A$. In this case the current is equal to $10\mu A$ at input voltage V_{in} equal to -1V.

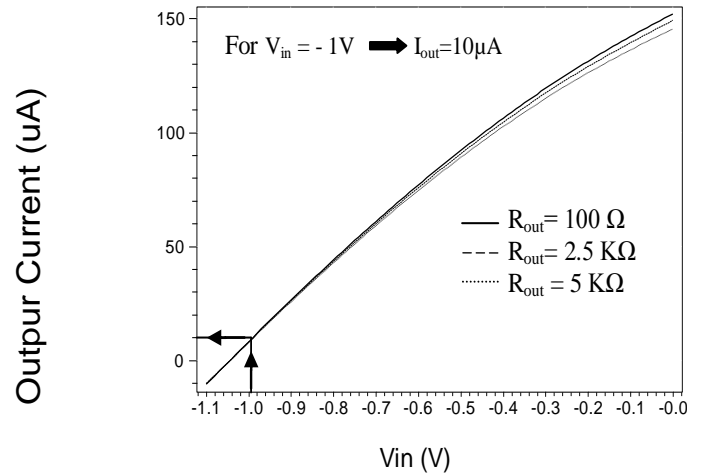


Figure 10. DC characteristics of V-I converter for different values of resistance R_{out}

From figure 8, the Miller operational amplifier is presented. Here in the polarisation circuit, an amplifier between the mirror's input and output transistor is necessary to achieve high current copy accuracy. In the other hand we use a simple amplifier composed of only two transistors. This amplifier is proposed by K. Tanno [19] to provide low voltage, low consumption and high performance as shown in figure 11.

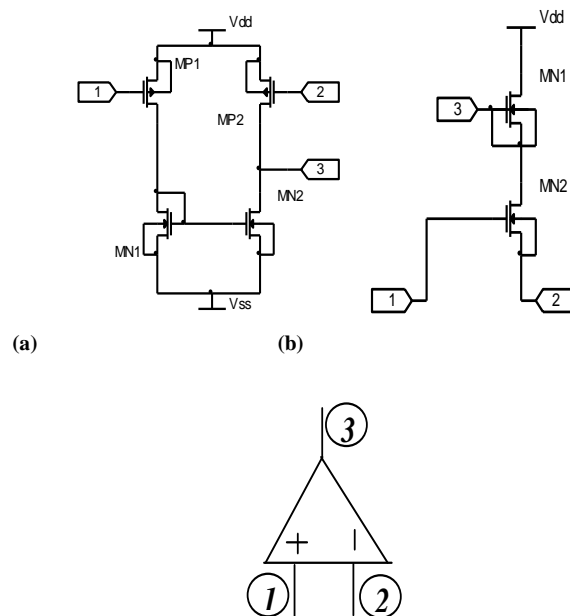


Figure 11. amplifier configuration:
(a) Simple differential amplifier structure [18]
(b) Two-transistor amplifier structure [19]

TABLE IV. COMPARISON TABLE OF MOST POPULAR DESIGNS WITH CURRENT WORK (*)

Resolution	OSR	SNR (dB)	Speed (MHz)	Power (mW)	Process (CMOS)	Signal Band width	Order	Ref. No.
14-bit	166	-	5.312	0.5	0.35µm	16 KHz	2	8
15-bit	16	96	40	44	0.25µm	1.25 MHz	2	9
10-bit	128	68	1	0.4	0.18µm	1 MHz	2	10
16-bit	64	93	3.2	5	0.35µm	25KHz	4	11
14-bit	24	85	2.2	200	0.35µm	100 KHz	6	12
11-bit	10	62.5	300	70	0.13µm	15 MHz	4	13
14-bit	96	85	53	15	0.18µm	300 KHz	2	14
16-bit	128	-	5.12	2.6	0.18µm	20 KHz	3	15
8-bit	64	49.7	1.024	6.6	0.6µm	8 KHz	1	16
14-bit*	88	85	14.08	9.8	0.35µm	80 KHz	2	This work

IV. RESULTS AND COMPARISON

The second order modulator is designed by using AMS technology 0.35µm CMOS process, the over sampling ratio is 88 with a signal band width of about 80 kHz.

All main parameters of the described modulator are summed up in Table III.

TABLE III. T DESIGNED MODULATOR PARAMETERS

Parameters	Value
Technology	AMS 0.35 µm
Order of modulator	2
Sampling Frequency (clock)	14.08 MHz
Signal Band width	80 KHz
Over sample Ratio(OSR)	88
References	±1V
Maximum Input	1 Vpp
Supply voltage	±1.5V
Resolution	14-bit
Signal to Noise Ratio (SNR)	85 dB
Dynamic range (DR)	86 dB
Quantizer resolution	1 bit
Power consumption	9.8 mW

The current state of the art in the design of $\Sigma\Delta$ modulator is limited by the technology and the sampling speeds it is able to achieve. Here is a comparison in table IV of the most popular designs which also compares the published works with the current work. It can be seen that the current work consumes less power than most published work and achieves the resolution of 14 bits using one of the technology AMS 0.35µm CMOS process.

V. CONCLUSION

The low-power-consumption modulator is designed with switched-capacitor techniques, and the resolution reaches 14 bits in AMS 0.35 µm CMOS process. Compared to other $\Sigma\Delta$ modulators, the second order single $\Sigma\Delta$ modulator has many advantages on performance, stability, area and system specification requirements, especially the power consumption.

When the power supply is ±1.5 V, the power consumption is only 9.8 mW.

In the future work, we will study and design of the different blocks constituting complete analogue to digital converter

(ADC) in transistor level, which is composed of an analog Sigma Delta modulator with a digital filter, in the other hand we will discuss about multi-bit discrete-time Sigma-Delta ADC.

REFERENCES

- [1] Delta Sigma Data Converters: Theory, Design and Simulation. IEEE Press, IEEE Circuits and Systems Society, 1997.
- [2] B. E. Boser and B. A. Wooley. The design of sigma-delta modulation and analog-to-digital converters. IEEE Journal of Solid-State Circuits, 23:1298–1308, December 1988.
- [3] Oversampling Delta-Sigma Converters. IEEE Press, 1992.
- [4] Top-Down Design of High-Performance Sigma-Delta Modulators. Kluwer Academic Publishers, 1999
- [5] Vineeta Upadhyay and Aditi Patwa, "Design Of First Order And Second Order Sigma Delta Analog To Digital Converter", International Journal of Advances in Engineering & Technology, July 2012.
- [6] Medeiro F., del Rio R., de la Rosa J.M., Pérez-Verdú B., A Sigma-Delta modulator design exemple : from specs to measurements, Baeceleonea, May 6-10, 2002
- [7] David Johns and Kenneth W. Martin : "Analog Integrated Circuit Design" John Wiley & Sons, 1997.
- [8] F. Munoz, A. P. VegaLeal, R. G. Carvajal, A. Torralba, J. Tombs, J. Ramirez-Angulo, "A 1.1V Low-Power $\Sigma\Delta$ Modulator For 14-bit 16KHz A/D Conversion"; The 2001 IEEE International Symposium on Circuits and Systems, Vol. 1, 6-9 May 2001
- [9] KiYoung Nam, Sang-Min Lee, David K. Su, and Bruce A. Wooley : " Voltage Low-Power Sigma-Delta Modulator for Broadband Analog-to-Digital Conversion" IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 40, NO. 9, SEPTEMBER 2005
- [10] H. Lee, C. Hsu, S. Huang, Y. Shih, C. Luo : " Designing low power of Sigma Delta Modulator for Biomedical Application", Biomedical Engineering-Applications, Basis & Communications, Vol. 17 No. 4, August 2005.
- [11] Hsin-Liang Chen, Yi-Sheng Lee, and Jen-Shiun Chiang: "Low Power Sigma Delta Modulator with Dynamic Biasing for Audio applications", Circuits and Systems, 5-8 Aug 2007. MWSCAS 2007. 50th Midwest Symposium on
- [12] Morizio J, Hoke I M, Kocak T, Geddie C, Hughes C, Perry J, Madhavapeddi S, Hood M, Lynch G, Kondoh H, Kumamoto T, Okuda T, Noda H, Ishiwaki M, Miki T, Nakaya M, "14-bit 2.2-MS/s sigma-delta ADC's", Solid-State Circuits, IEEE Journal of, Volume 35, Issue 7, July 2000 Page(s):968 – 976.
- [13] Di Giandomenico A, Paton S, Wiesbauer A, Hernandez L, Potscher T, Dorrer L, "A 15 MHz bandwidth sigma-delta ADC with 11 bits of resolution in 0.13/µm CMOS", Solid-State Circuits, IEEE Journal of, Volume 39, Issue 7, July 2004, Page(s): 1056- 1063.

- [14] Gaggli R, Wiesbauer A, Fritz G, Schranz Ch., Pessl P, "A 85-dB Dynamic Range Multibit Delta-Sigma ADC for ADSL-CO Applications in 0.18 μ m CMOS", *Solid-State Circuits, IEEE Journal of*, Volume 38, Issue 7, July 2003,Page(s): 1105 – 1115
- [15] Chen Yueyang, Zhong Shun'an, Dang Hua, "Design of A low-power-consumption and high-performance sigma-delta modulator", 2009 World Congress on Computer Science and Information Engineering
- [16] Boujelben S, Rebai Ch., Dallet D, Marchegay Ph., "Design and implementation of an audio analog to digital converter using oversamplingtechniques". 2001 IEEE
- [17] Laajimi Radwene; Gueddah Nawfil; Masmoudi Mohamed : "Low Power Variable Gain Amplifier with Bandwidth Of 80-300 MHz Using For Sigma-Delta Analogue to Digital Converter in Wireless Sensor Receiver" *International Journal of Computer Applications*;Apr2012, Vol. 43
- [18] Ahmed Nader Mohieldin, Edgar Sánchez-Sinencio, and José Silva-Martinez : ' Nonlinear effects in pseudo differential OTAs with CMFB', *IEEE Transactions On Circuits and Systems-II: Analog and Digital Signal Processing*, Vol. 50, No. 10, October 2003.
- [19] K. tanno, O. Ishizuka and Z. Tang : ' Low voltage and low frequency current mirror using a two- MOS subthreshold op-amp', *Electronics Letters* 28th March 1996 Vol. 32 No. 7

Short Answer Grading Using String Similarity And Corpus-Based Similarity

Wael H. Gomaa

Computer Science Department
Modern Academy for Computer Science & Management
Technology, Cairo, Egypt

Aly A. Fahmy

Computer Science Department
Faculty of Computers and Information, Cairo University
Cairo, Egypt

Abstract—Most automatic scoring systems use pattern based that requires a lot of hard and tedious work. These systems work in a supervised manner where predefined patterns and scoring rules are generated. This paper presents a different unsupervised approach which deals with students' answers holistically using text to text similarity. Different String-based and Corpus-based similarity measures were tested separately and then combined to achieve a maximum correlation value of 0.504. The achieved correlation is the best value achieved for unsupervised approach Bag of Words (BOW) when compared to previous work.

Keywords-Automatic Scoring; Short Answer Grading; Semantic Similarity; String Similarity; Corpus-Based Similarity.

I. INTRODUCTION

Educational community is growing endlessly with a growing number of students, curriculums and exams. Such a growing community raised the need for scoring systems that ease the burden of scoring numerous numbers of exams and in same time guarantees the fairness of the scoring process. Automatic Scoring (AS) systems evaluate student's answer by comparing it to model answer(s). The higher correlation between student and model answers the more efficient is the scoring system. The variety in curriculums forced the AS technology to handle different kinds of students' responses, such as writing, speaking and mathematics. Writing assessment comes in two forms Automatic Essay Scoring (AES) and Short Answer Grading. Speaking assessment includes low and high entropy spoken responses while mathematical assessments include textual, numeric or graphical responses. Design and implementation of Automatic Scoring system for questions as Multiple Choice, True-False, Matching and Fill in the blank is an easy task. AS systems designed for scoring essay questions is a more difficult and complicated task as student's answers require text understanding and analysis. This paper is concerned with the automatic scoring for answers for essay questions. This research presents an unsupervised approach that deals with student's answers holistically and uses text to text similarity measures [1, 2]. The proposed model calculates the automatic score by measuring the text similarity between each word in model answer to all words in the student's answer which saves the time spent by experts to generate predefined patterns and scoring rules.

Two types of text similarity measures are presented in this research, String-based similarity, and Corpus-based similarity. String-based similarity measures operate on string sequences

and character composition. Corpus-based works to identify the degree of semantic similarity between words; it depends on information derived from large corpora [3].

This paper is organized as follows:

- Section II presents related work of the main automatic short answer grading systems.
- Section III introduces the two main categories of used Similarity Algorithms.
- Section IV presents the used Data Set .
- Section V describes the proposed answer grading system.
- Section VI shows experiments' results.
- Finally, section VII presents conclusion.

II. RELATED WORK

This section describes the most famous short answer grading systems implemented for English language: C-rater [4, 5, 6], Oxford-UCLES [7, 8], Automark [9], IndusMarker [10], and Text-to-Text system [1, 2].

C-rater is the system developed by ETS, and is very reputational for high scoring accuracy for short answer responses. The reason behind high accuracy is using deep natural language processing to determine the relatedness of student response to the concepts listed in the rubric for an item. The C-rater engine applies a sequence of natural language processing steps including correcting students' spelling, determining the grammatical structure of each sentence, resolving pronoun reference, and analyzing paraphrases in the student responses [5, 6]. It has been validated on responses from multiple testing programs with different content areas including science, reading comprehension and history.

Oxford-UCLES is an information extraction short-answer scoring system that was developed at Oxford University. It uses pattern matching to evaluate the student's answers where patterns are discovered by human experts. First, it applies simple IE techniques as the nearest neighbors classification [7], then the machine learning methods like decision tree learning, Bayesian learning and inductive logic programming [8] are used.

Automarkis another system that uses IE techniques to explore the meaning or concept of text. The marking process depends mainly on content analysis in addition to specific

style features. Marking goes through 5 stages, they are discovering mark scheme templates, syntactic preprocessing, sentence analysis, pattern matching, and feedback module [9].

Indus Marker is a system that works on the structure of students' answer. It simply uses question answer markup language (QAML) to represent the required answer structures. The evaluation process starts with spell checking and some basic linguistic analysis, then the system matches the student's answer text structure with the required saved structure to compute the final mark [10].

Text-to-Text system as shown from the name; this system depends mainly on text comparison between student's answer and model answer. It doesn't use any predefined concepts or scoring rules in the evaluation process. In this approach, the evaluation process doesn't pay much attention to the subject materials, the student's answer methodology, the question type, the length of the answer and all such factors. Different semantic similarity measures were compared in [1, 2], including Knowledge-based and Corpus-based algorithms. This research depends on this approach by combining several String-based and Corpus-based similarity methods.

III. TEXT SIMILARITY MEASURES

Two categories of similarity algorithms are introduced; String-based and Corpus-based similarity. This section will handle the two measures in brief.

A. String-Based Similarity

String similarity measures operate on string sequences and character composition. A string metric is a metric that measures similarity or dissimilarity (distance) between two text strings for approximate string matching or comparison. Applying the concept of string metric; 13 algorithms of text similarity using Sim Metrics [11] are implemented. Six of them are character-based while the other seven are term-based distance measures

1) Character-based distance measures

Damerau-Levenshtein distance is a distance between two strings, given by counting the minimum number of operations needed to transform one string into the other, where an operation is defined as an insertion, deletion, or substitution of a single character, or a transposition of two adjacent characters [12,13].

Jaro algorithm is based on the number and order of the common characters between two strings; it takes into account typical spelling deviations and mainly used in the area of record linkage. [14, 15].

Jaro-Winkler distance is an extension of Jaro distance; it uses a prefix scale which gives more favorable ratings to strings that match from the beginning for a set prefix length [16].

Needleman-Wunsch algorithm is an example of dynamic programming, and was the first application of dynamic programming to biological sequence comparison. It performs a global to find the best alignment over the entire of two sequences. It is Suitable when the two sequences are of

similar length, with a significant degree of similarity throughout [17].

Smith-Waterman algorithm is an example of dynamic programming; it performs a local alignment to find the best alignment over the conserved domain of two sequences. It is useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context [18].

N-gram is a sub-sequence of n items from a given sequence of text; N-gram similarity algorithms compare the n-grams from each character or word in two strings. Distance is computed by dividing the number of similar n-grams by maximal number of n-grams [19].

2) Term-based distance measures

Block Distance is also known as Manhattan distance, boxcar distance, absolute value distance, L1 distance, city block distance and Manhattan distance, it computes the distance that would be traveled to get from one data point to the other if a grid-like path is followed. The Block distance between two items is the sum of the differences of their corresponding components [20].

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them.

Dice's coefficient is defined as twice the number of common terms in the compared strings divided by the total number of terms in both strings [21].

Euclidean distance or L2 distance is the square root of the sum of squared differences between corresponding elements of the two vectors.

Jaccard similarity is computed as the number of shared terms over the number of all unique terms in both strings [22].

Matching Coefficient is a very simple vector based approach which simply counts the number of similar terms, (dimensions), on which both vectors are non-zero.

Overlap coefficient is similar to the Dice's coefficient, but considers two strings a full match if one is a subset of the other.

B. Corpus-Based Similarity

Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora.

Latent Semantic Analysis (LSA) [23] is the most popular technique of Corpus-Based Similarity, LSA assumes that words that are close in meaning will occur in similar pieces of text. A matrix containing word counts per paragraph (rows represent unique words and columns represent each paragraph) is constructed from a large piece of text and a mathematical technique which called singular value decomposition (SVD) is used to reduce the number of columns while preserving the similarity structure among rows. Words are then compared by taking the cosine of the angle between the two vectors formed by any two rows.

Explicit Semantic Analysis (ESA) [24] is a measure used to compute the semantic relatedness between two arbitrary texts. The Wikipedia-Based technique represents terms (or texts) as high-dimensional vectors, each vector entry presenting the TF-IDF weight between the term and one Wikipedia article. The semantic relatedness between two terms (or texts) is expressed by the cosine measure between the corresponding vectors.

Pointwise Mutual Information - Information Retrieval (PMI-IR) [25] is a method for computing the similarity between pairs of words, it uses AltaVista's Advanced Search query syntax to calculate probabilities. The more often two words co-occur near each other on a web page, the higher is their PMI-IR similarity score.

Extracting DIS tributionally similar words using CO-occurrences (DISCO¹) [26, 27] Distributional similarity between words assumes that words with similar meaning occur in similar context. Large text collections are statistically analyzed to get the distributional similarity. DISCO is a method that computes distributional similarity between words by using a simple context window of size ± 3 words for counting co-occurrences. When two words are subjected for exact similarity DISCO simply retrieves their word vectors from the indexed data, and computes the similarity according to Lin measure [28]. If the most distributionally similar word is required; DISCO returns the second order word vector for the given word.

DISCO has two main similarity measures DISCO1 and DISCO2:

- DISCO1: Computes the first order similarity between two input words based on their collocation sets.
- DISCO2: Computes the second order similarity between two input words based on their sets of distributionally similar words.

This research, handled the corpus-based approach via DISCO using the two main similarity measures DISCO1 and DISCO2.

IV. THE DATA SET

Texas² short answer grading data set is used [2]. It consists of ten assignments between four and seven questions each and two exams with ten questions each. These assignments/exams were assigned to an introductory computer science class at the University of North Texas. The assignments were administered as part of a Data Structures course at the University of North Texas. For each assignment, the student answers were collected via an online learning environment.

The data set as a whole contains 80 questions and 2273 student answers. The answers were scored by two human judges, using marks between 0 (completely incorrect) and 5 (perfect answer). Data set creators treated the average grade of the two evaluators as the gold standard to examine the automatic scoring task.

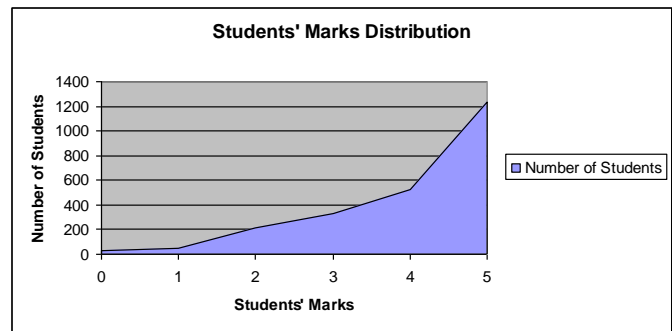


Figure 1. Students' Marks Distribution

Table I. Sample Questions, Model Answer and Students Answers

	Question, Model Answer and Student answers	Average Grades
Question :	What is a variable?	
Model Answer :	A location in memory that can store a value.	
Student answer 1:	A variable is a location in memory where a value can be stored.	5
Student answer 2:	A named object that can hold a numerical or letter value.	4
Student answer 3:	Variable can be a integer or a string in a program	2
Question :	What is the role of a header-file?	
Model Answer :	To store a class interface, including data members and member function prototypes.	
Student answer 1:	a header file is a file used to store a list of prototype functions and data members.	5
Student answer 2:	to declare the functions being used in the classes.	3
Student answer 3:	Header files have reusable source code in a file that a programmer can use.	2.5

Figure 1 shows the students' marks distribution and table I represents a sample question, model answer, student answers and average grade.

V. ANSWER GRADING SYSTEM

Similar to all systems of automatic short answer grading, this system is based on measuring the similarity between the student's answer and the model answer to produce the final score. Then Pearson's correlation coefficient is used to specify the correlation between automatic score and average human grades.

The system goes through three stages:

The First stage is measuring the similarity between model answer and student answer using 13 String-Based algorithms previously described in section III. In this stage four methods are used to deal with strings in model and students answer; Raw, Stop, Stem, StopStem. The similarity in Raw method is computed without applying any Natural Language Processing (NLP) task. Stop Words Removing is applied in the Stop method using stop list that contains 429 words. In Stem method Porter Stemmer [29] is used to replace each non-stop with its stem without removing the stop words. Both Stop Words Removing and Stemming tasks are applied in StopStem method. Table II represents a sample of student answer using 4 methods.

¹http://www.linguatools.de/disco/disco_en.html

²<http://lit.csci.unt.edu/index.php?P=research/downloads>

Table II. Sample Student Answer with 4 Forms

Method	Student Answer
Raw	removing logical errors testing for valid data random data and actual data
Stop	removing logical errors testing valid data random data actual data
Stem	remov logic error test for valid data random data and actual data
StopStem	remov logic error test valid data random data actual data

The Second stage is measuring the similarity using DICSO1 and DISCO2 corpus-based similarity. In this stage three tasks are performed; Removing the stop words, Getting distinct words and Constructing the similarity matrix. The similarity matrix represents the similarity between each distinct word in the model answer and each distinct word in the student's answer. Each row represents one word in the model answer, and each column represents one word in the student's answer. The last two columns represent the maximum and the average similarity of each word in the model answer.

After constructing the similarity matrix, the final overall similarity is computed with two methods - Max Overall Similarity and Average Overall Similarity - by computing the average of the last two columns (Max, Average).The final overall similarity refers to the student's mark. For more clarification consider the following walkthrough example to

Table III. Similarity matrix using DISCO2 Corpus-based similarity

	Header	File	reusable	Source	code	Programmer	MAX	AVG
Store	0.282	0.399	0.12	0.285	0.266	0.193	0.399	0.257
Class	0.035	0.052	0.044	0.049	0.067	0.031	0.067	0.046
Interface	0.439	0.697	0.234	0.383	0.522	0.389	0.697	0.444
Including	0.003	0.009	0.004	0.009	0.005	0.008	0.009	0.006
Data	0.468	0.61	0.163	0.017	0.45	0.253	0.61	0.326
Member	0.026	0.037	0.006	0.095	0.046	0.132	0.132	0.057
Function	0.2	0.261	0.057	0.3	0.285	0.147	0.3	0.208
prototypes	0.078	0.106	0.139	0.125	0.019	0.094	0.139	0.093
Final Overall Similarity							0.294	0.179

VI. EXPERIMENTS' RESULTS AND DISCUSSION

Pearson's correlation coefficient measure was used to specify the correlation between automatic score and average human grades.

A. Experiments Results using String-Based Similarity

As mentioned above; 13 string-based algorithms were tested with four different methods Raw, Stop, Stem and Stop Stem. Table IV represents the correlation results between model and student answer using both Character-based and Term-based measures. In character-based distance measures; N-gram similarity got the best correlation value 0.435 applied to the raw text by mixing the results obtained from both bi-gram and tri-gram similarity measures.

measure the similarity between the following Model and student answer:

- Model Answer: "To store a class interface including data members and member function prototypes."
- Student Answer: "Header files have reusable source code in a file that a programmer can use."

First step is removing the stop words from the two strings ("To, a, and" in model answer and "have, in, a, that, can, use" in student answer).

Second step is getting the distinct words from the two strings; two words are considered equal if they have the same stem ("members, member" in model answer and "files, file in student answer").

Third step is constructing the similarity matrix. Table III represents the similarity matrix using DISCO2 with Wikipedia data packets.

The Third stage is combining the similarity values obtained from both string-based and corpus-based measures. Many researches adopted the idea of mixing the results from different measures to enhance the overall similarity [30, 31, 32, and 33]. The steps of the proposed combining task are illustrated in the next section.

In Term-based distance measures; Block Distance got the best correlation value 0.382 applied to the text after removing the stop words from both model and student answers.

Stop word removing task enhanced the correlation results especially in Term-based measures. Stemming process didn't enhance the results for all cases.

B. Experiments Results using Corpus-Based Similarity

Disco measures are applied by using two data packets - Wikipedia and British National Corpus (BNC); features of both are presented in table V.

Table IV. Similarity matrix using DISCO2 Corpus-based similarity

	Raw	Stop	Stem	StopStem
Character-based distance measures				
Damerau-Levenshtein	0.338	0.324	0.317	0.315
Jaro	0.144	0.229	0.146	0.205
Jaro-Winkler	0.151	0.245	0.169	0.223
Needleman-Wunsch	0.265	0.265	0.255	0.258
Smith-Waterman	0.361	0.341	0.351	0.331
N-gram (bi-gram+tri-gram)	0.435	0.416	0.413	0.398
Term-based distance measures				
Block Distance	0.375	0.382	0.34	0.291
Cosine similarity	0.376	0.377	0.344	0.308
Dice's coefficient	0.368	0.379	0.337	0.307
Euclidean distance	0.326	0.338	0.312	0.281
Jaccard similarity	0.332	0.349	0.311	0.294
Matching Coefficient	0.305	0.339	0.294	0.264
Overlap coefficient	0.374	0.368	0.336	0.286

Table V. Disco Data Packets

	Wikipedia	BNC
Packet Name	en-wikipedia-20080101	en-BNC-20080721
Packet Size	5.9 Gigabyte	1.7 Gigabyte
Number of Tokens	267 million	119 million
Number of queriable words	220,000	122,000

Table VI represents the correlation results between all the model and student answer using Disco1 and Disco2 measures. As mentioned in section V, MAX and AVG refer to Max Overall Similarity and Average Overall Similarity respectively.

Table VI. Disco Data Packets

	Wikipedia		BNC	
	MAX	AVG	MAX	AVG
Disco1	0.465	0.445	0.450	0.412
Disco2	0.428	0.410	0.415	0.409

In Corpus-based measures; Disco1 similarity got the best correlation value 0.475 using Wikipedia data packet and Max overall similarity method. Similarity measures using Wikipedia packet got higher correlation than BNC due to the role of the corpus size and other features shown in table V. Using Max overall similarity method clearly enhanced the correlation results in all cases in corpus-based measures.

C. Experiments Results via Combining String-Based and Corpus-Based similarity

As previously mentioned, many researches adopt the idea of mixing results from different measures to enhance the overall similarity. The proposed system combined the best algorithm for each category. The three selected measures are:

- N-gram represents character-based string similarity and is applied to raw text.

- Block Distance represents term-based string similarity and is applied to text after removing stop words.
- Disco1 using Max overall similarity which represents corpus-based similarity.

Similarity values resulting from the three measures are compared, the max and average similarity value for each student's answer are selected, and then the correlation between all students and model answers is recomputed.

The four possible combinations are represented in table VII. These cases emphasize the idea of mixing String-Based Similarity measures with the Corpus-based similarity measures to get the advantages of both. The correlation results are enhanced from the best value achieved from applying all the measures separately 0.465 to 0.504 resulting from combining N-gram and Disco1 measures.

Table VII. Correlation Results based on combining method

N-gram	Block Distance	Disco1	MAX	AVG
✗	✓	✓	0.457	0.414
✓	✗	✓	0.504	0.470
✓	✓	✗	0.411	0.394
✓	✓	✓	0.475	0.443

D. Discussion

As previously mentioned in section II; the most related research to this work was introduced in [1, 2]. The discussion here is a comparison between results from the previously related researches and results from the proposed research. Care is given to researches that deal with text as bag of words (BOW) in unsupervised way, where neither complex NLP tasks nor machine learning algorithms were applied. The dataset experimented in [1] contained 21 questions and 610 answers. LSA (BNC), LSA (Wikipedia), ESA (Wikipedia) and tf*idf were experimented. The results were 0.407, 0.428, 0.468 and 0.364 respectively. The dataset experimented in [2] was Texas dataset previously introduced in section IV .LSA, ESA and tf*idf were experimented and results were 0.328, 0.395 and 0.281 respectively.

Compared to previous results the proposed system achieved better results in most experimented cases. String-based similarity measures enhanced the correlation results if compared to simple tf*idf method. Also Disco similarity measures achieved better results than the most known Corpus-based methods LSA and ESA.

Combining String-based and Corpus-based similarity in unsupervised way raised the correlation results to 0.504. This value is the best correlation result achieved compared to other previous work and is very promising as these measures don't need any complex supervised learning task or NLP tasks such as Part of Speech and Syntax parsing. Also this value is very near to correlation values obtained from learning using different supervised machine learning algorithms and graph alignment in [2].

VII. CONCLUSION

In this research, short answer grading task is handled from an unsupervised approach which is bag of words. This approach is easy to implement as it neither requires complex NLP tasks nor supervised learning algorithms. The used data set contains 81 questions and 2273 student answers.

The proposed model goes through three stages: The First stage is measuring the similarity between model answer and student answer using 13 String-Based algorithms. Six of them were Character-based and the other seven were Term-based measures. The best correlation values achieved using Character-based and term-based were 0.435 and 0.382 using N-gram and Block distance respectively. The Second stage was measuring the similarity using DICS01 and DISCO2 Corpus-based similarity. Disco1 achieved 0.465 correlation value using the max overall similarity.

The Third stage was measuring the similarity by combing String-based and Corpus-based measures. The best correlation value 0.504 was obtained from mixing N-gram with Disco1 similarity values. Proposed model achieved great results compared to previous works. The near future work focus on applying short answer grading to other language like Arabic. A very encouraging factor is the ability of Disco package to work with nine languages. A main obstacle for this task is the unavailability of short answer grading data sets in other language than English language.

REFERENCES

- [1] M. Mohler and R. Mihalcea, "Text-to-text semantic similarity for automatic short answer grading", In Proceedings of the European Association for Computational Linguistics (EACL 2009), Athens, Greece, 2009.
- [2] M. Mohler, R. Bunescu & R. Mihalcea, "Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics, pp. 752-762, 2011.
- [3] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based approaches to text semantic similarity", In Proceedings of the American Association for Artificial Intelligence (AAAI 2006), Boston, 2006.
- [4] C. Leacock and M. Chodorow, "C-rater: Automated Scoring of Short-Answer Question", Computers and the Humanities, vol. 37, no. 4, pp. 389-405, Nov. 2003.
- [5] J. Sukkarieh and S. Stoyanchev, "Automating model building in C-rater", In Proceedings of the Workshop on Applied Textual Inference, pages 6169, Suntec, Singapore, August, 2009.
- [6] J. Z. Sukkarieh & J. Blackmore, "c-rater: Automatic Content Scoring for Short Constructed Responses", Proceedings of the 22nd International FLAIRS Conference, Association for the Advancement of Artificial Intelligence, 2009.
- [7] J.Z. Sukkarieh, S.G. Pulman, and N. Raikes, "Auto-Marking 2: An Update on the UCLES-Oxford University research into using Computational Linguistics to Score Short, Free Text Responses", International Association of Educational Assessment, Philadelphia, 2004.
- [8] J. Z. Sukkarieh and S. G. Pulman, "Automatic Short Answer Marking". Proceedings of the 2nd Workshop on Building Educational Applications Using NLP, Association for Computational Linguistics, pp. 9-16, June 2005.
- [9] T. Mitchell, T. Russel, P. Broomhead and N. Aldridge, "Towards robust computerized marking of free-text responses". Proceedings of the Sixth

- International Computer Assisted Assessment Conference, Loughborough, UK: Loughborough University, 2002.
- [10] Raheel Siddiqi, Christopher J. Harrison, and Rosheena Siddiqi, "Improving Teaching and Learning through Automated Short-Answer Marking" IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 3, NO. 3, JULY-SEPTEMBER, 2010.
- [11] S. Chapman, "Simmetrics: a java & c# .net library of similarity metrics", <http://sourceforge.net/projects/simmetrics/>, 2006.
- [12] P. A. V. Hall and G. R. Dowling, "Approximate string matching, Comput. Surveys", 12:381-402, 1980.
- [13] J. L. Peterson, "Computer programs for detecting and correcting spelling errors", Comm. Assoc. Comput. Mach., 23:676-687, 1980.
- [14] M. A. Jaro, "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida". Journal of the American Statistical Society 84 (406): 414-20, 1989.
- [15] M. A. Jaro, "Probabilistic linkage of large public health data file", Statistics in Medicine 14 (5-7), 491-8, 1995.
- [16] W. E. Winkler, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage", Proceedings of the Section on Survey Research Methods (American Statistical Association): 354-359, 1990.
- [17] Needleman, B.Saul, and Wunsch, D. Christian, "A general method applicable to the search for similarities in the amino acid sequence of two proteins", Journal of Molecular Biology 48(3): 443-53, 1970.
- [18] Smith, F. Temple, Waterman, S. Michael, "Identification of Common Molecular Subsequences". Journal of Molecular Biology 147: 195-197, 1981.
- [19] Alberto Barr'on-Cedeño, Paolo Rosso, Eneko Agirre, and Gorka Labaka, "Plagiarism Detection across Distant Language Pairs", In Proceedings of the 23rd International Conference on Computational Linguistics, pages 37-45, 2010.
- [20] Eugene F. Krause, "Taxicab Geometry", Dover. ISBN 0-486-25202-7, 1987.
- [21] L. Dice, "Measures of the amount of ecologic association between species", Ecology, 26(3), 1945.
- [22] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura", Bulletin de la Société Vaudoise des Sciences Naturelles 37, 547-579, 1901.
- [23] T.K. Landauer and S.T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge", Psychological Review, 104, 1997.
- [24] E. Gabrilovich and S. Markovitch, "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis", Proceedings of the 20th International Joint Conference on Artificial Intelligence, pages 6-12, 2007.
- [25] P. Turney, "Mining the web for synonyms: PMI-IR versus LSA on TOEFL", In Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001), 2001.
- [26] Peter Kolb, "Experiments on the difference between semantic similarity and relatedness", In Proceedings of the 17th Nordic Conference on Computational Linguistics - NODALIDA '09, Odense, Denmark, May 2009.
- [27] Peter Kolb, "DISCO: A Multilingual Database of Distributionally Similar Words", In Proceedings of KONVENS-2008, Berlin, 2008.
- [28] D. Lin, "Extracting Collocations from Text Corpora. In Workshop on Computational Terminology", 57-63, Montreal, Canada, 1998.
- [29] M. F. Porter, "An algorithm for suffix stripping", Program, 14, 130, 1980.
- [30] Y. Li, D. McLean, Z. Bandar, J. O'Shea, K. Crockett, "Sentence similarity based on semantic nets and corpus statistics", IEEE Transactions on Knowledge and Data Engineering, 18(8), 1138-1149, 2006.
- [31] A. Islam, D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity", ACM Transactions on Knowledge Discovery from Data, 2(2), 1-25, 2008.
- [32] Nitish Aggarwal, Kartik Asooja, Paul Buitelaar. "DERI&UPM: Pushing Corpus Based Relatedness to Similarity: Shared Task System

Description", First Joint Conference on Lexical and Computational Semantics (*SEM), pages 643–647, Montreal, Canada, Association for Computational Linguistics, June 7-8, 2012.

- [33] Davide Buscaldi, Ronan Tournier, Nathalie Aussenac-Gilles and Josiane Mothe, "IRIT: Textual Similarity Combining Conceptual Similarity with an N-Gram Comparison Method", First Joint Conference on Lexical and Computational Semantics (*SEM), pages 552–556, Montreal, Canada, Association for Computational Linguistics, June 7-8, 2012.

AUTHORS PROFILE



Wael Hasan Gomaa, is currently working as a teacher assistant, Computers Science department, Modern Academy for Computer Science & Management Technology, Cairo, Egypt. He is a Ph.D student, Faculty of Computer and Information, Cairo University, Egypt in the field of Automatic Assessment under supervision of Prof. Aly Aly Fahmy. He received his B.Sc and Master degrees from Faculty of Computers and Information, Helwan University, Egypt. His master thesis was entitled "Text Mining Hybrid Approach for Clustered Semantic Analysis". His research interests include Natural Language Processing, Artificial Intelligence, Data Mining and Text Mining. E-mail: Wael.goma@gmail.com



Prof. Aly Aly Fahmy, is the former Dean of the Faculty of Computing and Information, Cairo University and a Professor of Artificial Intelligence and Machine Learning, in the department of Computer Science. He graduated from the Department of Computer Engineering, Technical College with honor degree. He specialized in Mathematical Logic and did his research with Dr. Hervey Gallaire, the former vice president of Xerox Global. He

received a master's degree from the National School of Space and Aeronautics ENSAE, Toulouse, France, 1976 in the field of Logical Data Base systems and then obtained his PhD from the Centre for Studies and Research – CERT-DERI, 1979, Toulouse - France in the field of Artificial Intelligence.

He received practical training in the field of Operating Systems and Knowledge Based Systems in Germany and the United States of America. He participated in several national projects including the establishment of the Egyptian Universities Network (currently hosted at the Egyptian Academy of Scientific Research at the Higher Ministry of Education), building Expert Systems in the field of iron and steel industry and building Decision Support Systems for many national entities.

Prof. Fahmy's main research areas are: Data and Text Mining, Mathematical Logic, Computational Linguistics, Text Understanding and Automatic Essay Scoring and Technologies of Man- Machine Interface in Arabic. He published many refereed papers and authored the book "Decision Support Systems and Intelligent Systems" in Arabic.

He was the Director of the first Center of Excellence in Egypt in the field of Data Mining and Computer Modeling (DMCM) in the period of 2005-2010. DMCM was a virtual research center with more than 40 researchers from universities and industry. It was founded by an initiative of Dr. Tarek Kamel, Minister of Communications and Information Technology.

Prof. Aly Fahmy is currently involved in the implementation of the exploration project of Master's and Doctorate theses of Cairo University with the supervision of Prof. Dr. Hussein Khaled, Vice President of Cairo University for Postgraduate Studies and Research. The project aims to assess the practical implementation of Cairo University late strategic research plan, and to assist in the formulation of the new strategic research plan for the coming 2011 - 2015. E-mail: Aly.Fahmy@cu.edu.eg

Hybrid intelligent system for Sale Forecasting using Delphi and adaptive Fuzzy Back-Propagation Neural Networks

Attariuas Hicham

LIST, Group Research in Computing and
Telecommunications (ERIT) FST
Tangier, BP : 416, Old Airport Road, Morocco

Bouhorma Mohammed, Sofi Anas

LIST, Group Research in Computing and
Telecommunications (ERIT) FST
Tangier, BP: 416, Old Airport Road, Morocco

Abstract—Sales forecasting is one of the most crucial issues addressed in business. Control and evaluation of future sales still seem concerned both researchers and policy makers and managers of companies. this research propose an intelligent hybrid sales forecasting system Delphi-FCBPN sales forecast based on Delphi Method, fuzzy clustering and Back-propagation (BP) Neural Networks with adaptive learning rate. The proposed model is constructed to integrate expert judgments, using Delphi method, in enhancing the model of FCBPN. Winter's Exponential Smoothing method will be utilized to take the trend effect into consideration. The data for this search come from an industrial company that manufactures packaging. Analyze of results show that the proposed model outperforms other three different forecasting models in MAPE and RMSE measures.

Keywords-Component; Hybrid intelligence approach; Delphi Method; Sales forecasting; fuzzy clustering; fuzzy system; back propagation network.

I. INTRODUCTION

Sales forecasting plays a very important role in business strategy. To strengthen the competitive advantage in a constantly changing environment, the manager of a company must make the right decision in the right time based on their formation at hand. Obtaining effective sales forecasts in advance can help decision makers to calculate production and material costs, determine also the sales price, strategic Operations Management, etc.

Hybrid intelligent system denotes system that utilizes a parallel combination of methods and techniques from artificial intelligence. As hybrid intelligent systems can solve non-linear prediction, this article proposes an integration of a hybrid system FCBPN within an architecture of ERP to improve and extend the management sales module to provide sales forecasts and meet the needs of decision makers of the company.

The remainder of the article is constructed as follows: Section 2 is the literature review. Section 3 describes the construction and role of each component of the proposed sale forecasting system Delphi-FCBPN Section 4 describes the sample selection and data analysis. Finally, section 5 provides a summary and conclusions.

II. LITERATURE AND RELATED RESEARCH

Enterprise Resource Planning (ERP) is a standard of a complete set of enterprise management system .It emphasizes integration of the flow of information relating to the major functions of the firm [29]. There are four typical modules of ERP, and the sales management is one of the most important modules. Sales management is highly relevant to today's business world; it directly impacts the working rate and quality of the enterprise and the quality of business management. Therefore, Integrate sales forecasting system with the module of the sales management has become an urgent project for many companies that implement ERP systems.

Many researchers conclude that the application of BPN is an effective method as a forecasting system, and can also be used to find the key factors for enterprisers to improve their logistics management level. Zhang, Wang and Chang (2009) [28] utilized Back Propagation neural networks (BPN) in order to forecast safety stock. Zhang, Haifeng and Huang (2010)[29] used BPN for Sales Forecasting Based on ERP System. They found out that BPN can be used as an ac-curate sales forecasting system.

Reliable prediction of sales becomes a vital task of business decision making. Companies that use accurate sales forecasting system earn important benefits. Sales forecasting is both necessary and difficult. It is necessary because it is the starting point of many tools for managing the business: production schedules, finance, marketing plans and budgeting, and promotion and advertising plan. It is difficult because it is out of reach regardless of the quality of the methods adopted to predict the future with certainty. Parameters are numerous, complex and often unquantifiable.

Recently, the combined intelligence technique using artificial neural networks (ANNs), fuzzy logic, Particle Swarm Optimization (PSO), and genetic algorithms (GAs) has been demonstrated to be an innovative forecasting approach. Since most sales data are non-linear in relation and complex, many studies tend to apply Hybrid models to time-series forecasting. Kuo and Chen (2004)[20] use a combination of neural networks and fuzzy systems to effectively deal with the marketing problem.

The rate of convergence of the traditional back-propagation networks is very slow because it's dependent upon the choice of value of the learning rate parameter. However, the experimental results (2009 [25]) showed that the use of an adaptive learning rate parameter during the training process can lead to much better results than the traditional neural network model (BPN).

Many papers indicate that the system which uses the hybridization of fuzzy logic and neural networks can more accurately perform than the conventional statistical method and single ANN. Kuo and Xue (1999) [21] proposed a fuzzy neural network (FNN) as a model for sales forecasting. They utilized fuzzy logic to extract the expert's fuzzy knowledge. Toly Chen (2003) [27] used a model for wafer fab prediction based on a fuzzy back propagation network (FBPN). The proposed system is constructed to incorporate production control expert judgments in enhancing the performance of an existing crisp back propagation network. The results showed the performance of the FBPN was better than that of the BPN. Efendigil, Önü, and Kahraman (2009) [16] utilized a forecasting system based on artificial neural networks ANNs and adaptive network based fuzzy inference systems (ANFIS) to predict the fuzzy demand with incomplete information.

Attariuas, Bouhorma and el Fallahi[30] propose hybrid sales forecasting system based on fuzzy clustering and Back-propagation (BP) Neural Networks with adaptive learning rate (FCBPN). The experimental results show that the proposed model outperforms the previous and traditional approaches (BPN, FNN, WES, KGFS). Therefore, it is a very promising solution for industrial forecasting.

Chang and Wang (2006)[6] used a fuzzy back-propagation network (FBPN) for sales forecasting. The opinions of sales managers about the importance of each input, were converted into prespecified fuzzy numbers to be integrated into a proposed system. They concluded that FBPN approach outperforms other traditional methods such as Grey Forecasting, Multiple Regression Analysis and back propagation networks.

Chang, Liu, and Wang (2006)[7] proposed a fusion of SOM, ANNs, GAs and FRBS for PCB sales forecasting. They found that performance of the model was superior to previous methods that proposed for PCB sales forecasting.

Chang, Wang and Liu (2007) [10] developed a weighted evolving fuzzy neural network (WEFuNN) model for PCB sales forecasting. The proposed model was based on combination of sales key factors selected using GRA. The experimental results that this hybrid system is better than previous hybrid models.

Chang and Liu (2008)[4] developed a hybrid model based on fusion of case-based reasoning (CBR) and fuzzy multicriteria decision making. The experimental results showed that performance of the fuzzy casebased reasoning (FCBR) model is superior to traditional statistical models and BPN.

A Hybrid Intelligent Clustering Forecasting System was proposed by Kyong and Han (2001)[22]. It was based on Change Point Detection and Artificial Neural Networks. The basic concept of proposed model is to obtain significant

intervals by change point detection. They found out that the proposed models are more accurate and convergent than the traditional neural network model (BPN).

Chang, Liu and Fan (2009) [5] developed a K-means clustering and fuzzy neural network (FNN) to estimate the future sales of PCB. They used K-means for clustering data in different clusters to be fed into independent FNN models. The experimental results show that the proposed approach outperforms other traditional forecasting models, such as, BPN, ANFIS and FNN.

Recently, some researchers have shown that the use of the hybridization between fuzzy logic and GAs leading to genetic fuzzy systems (GFSs) (Cordón, Herrera, Hoffmann, & Magdalena (2001) [13]) has more accurate and efficient results than the traditional intelligent systems. Casillas, & Martínez López (2009) [24], Martínez López & Casillas (2009) [23], utilized GFS in various case Management. They have all obtained good results.

Chang, Wang and Tsai (2005)[3] used Back Propagation neural networks (BPN) trained by a genetic algorithm (ENN) to estimate demand production of printed circuit board (PCB). The experimental results show that the performance of ENN is greater than BPN.

Hadavandi, Shavandi and Ghanbari (2011) [18] proposed a novel sales forecasting approach by the integration of genetic fuzzy systems (GFS) and data clustering to construct a sales forecasting expert system. They use GFS to extract the whole knowledge base of the fuzzy system for sales forecasting problems. Experimental results show that the proposed approach outperforms the other previous approaches.

This article proposes an intelligent hybrid sales forecasting system Delphi-FCBPN sales forecast based on Delphi Method, fuzzy clustering and Back-propagation (BP) Neural Networks with adaptive learning rate (FCBPN) for sales forecasting in packaging industry

III. DEVELOPMENT OF THE DELPHI-FCBPN MODEL

The proposed approach is composed of three stage as shown in Figure 1 : (1) Stage of data collection: collection of key factors that influence sales will be made using the Delphi method through experts judgments; (2) Stage of Data preprocessing: Use Rescaled Range Analysis (R/S) to evaluate the effects of trend. Winter's Exponential Smoothing method will be utilized to take the trend effect into consideration. (3) Stage of learning by FCBPN: We use hybrid sales forecasting system based on Delphi, fuzzy clustering and Back-propagation (BP) Neural Networks with adaptive learning rate (FCBPN).

The data for this study come from an industrial company that manufactures packaging in Tangier from 2001-2009. Amount of monthly sales is seen as an objective of the forecasting model.

A. Stage of data collection

Data collection will be implemented based on production-control expert judgments. Some sales managers and production control experts are requested to express their

opinions about the importance of each input parameter in predicting the sales, and then we apply the Delphi Method to select the key factors for forecasting packaging industry.

1) *The collection of factor that influence sales*

Packaging industry is an industry with highly variant environment. The variables of packaging industry sales can be subdivided in three domains: (1) the market demand domain; (2) macroeconomics domain; (3) industrial production domain. To collect all possible factors, which can affect the packaging industry sales from the three domains mentioned earlier, some sales managers and production control experts are requested to list all possible attributes affecting packaging industry sales

2) *Delphi Method to select the factors affecting sales*

The Delphi Method was first developed by Dalkey and Helmer (1963) in corporation and has been widely applied in many management areas, e.g. forecasting, public policy analysis, and project planning. The principle of the Delphi method is the submission of a group of experts in several rounds of questionnaires. After each round, a synthesis anonymous of response with experts' arguments is given to them. The experts were then asked to revise their earlier answers in light of these elements. It is usually found as a result of this process (which can be repeated several times if necessary), the differences fade and responses converge towards the "best" answer.

The Delphi Method was used to choose the main factors, which would influence the sales quantity from all possible that were collected in this research. The procedures of Delphi Method are listed as follows:

1. Collect all possible factors from ERP database, which may affect the monthly sales from the domain experts. This is the first questionnaire survey.
2. Conduct the second questionnaire and ask domain experts select assign a fuzzy number ranged from 1 to 5 to each factor. The number represents the significance to the sales.
3. Finalize the significance number of each factor in the questionnaire according to the index generated in step 3.

Repeat 2 to 3 until .the results of questionnaire converge

B. *Data preprocessing stage*

Based on Delphi method, the key factors that influence sales are (K1, K2, K3) (see Table 1): When the seasonal and trend variation is present in the time series data, the accuracy of forecasting will be influenced. R/S analysis will be utilized to detect if there is this kind of effects of serious data. If the effects are observed, Winter's exponential smoothing will be used to take the effects of seasonality and trend into consideration.

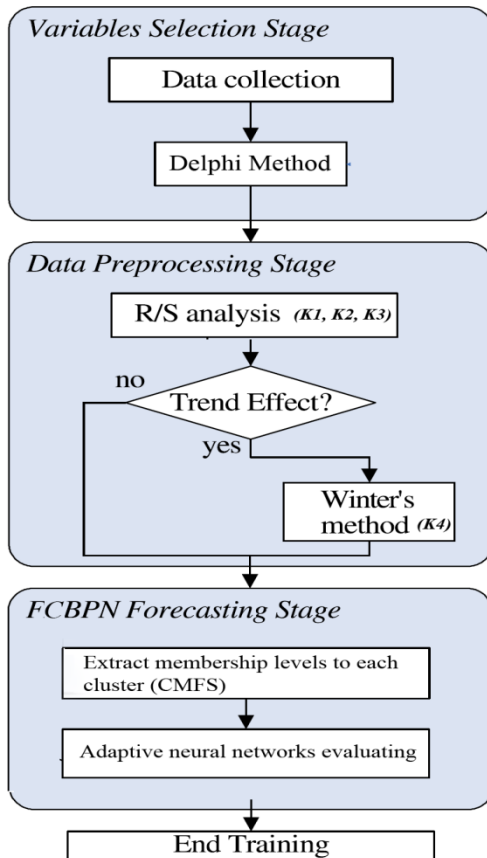


Figure 1: Architecture of the Delphi-FCBPN model

Input Description

K_1	Manufacturing consumer index
N_1	Normalized manufacturing consumer index
K_2	Offers competitive index
N_2	Normalized offers competitive index
K_3	packaging total production value Index
N_3	Normalized packaging total production value Index
K_4	Preprocessed historical data (WES)
N_4	Normalized preprocessed historical data (WES)
Y^0	Actual historical monthly packaging sales
Y	Normalized Actual historical monthly packaging sales

Table1: Description of input forecasting model.

1) *R/S analysis (rescaled range analysis)*

For eliminating possible trend influence, the rescaled range analysis, invented by Hurst (Hurst, Black, & Simaika, 1965), is used to study records in time or a series of observations in different time. Hurst spent his lifetime studying the Nile and the problems related to water storage. The problem is to determine the design of an ideal reservoir on the basis of the given record of observed discharges from the lake. The R/S analysis will be introduced as follows:

Consider the $XZ\{x_1, x_2, \dots, x_n\}$, x_i is the sales amount in period i , and compute M_N where

$$M_N = \frac{1}{N} \sum_{i=1}^N X_i$$

The standard deviation S is defined as

$$s = \sqrt{\frac{\sum_{i=1}^N (X_i - M_N)^2}{N}}$$

For each point i in the time series, we compute

$$X(t, N) = \sum_{i=1}^t X_i - M_N$$

$$R = \max_{1 \leq t \leq N} X(t, N) - \min_{1 \leq t \leq N} X(t, N)$$

We computed the H coefficient as

$$H = \text{Ln} \left(\frac{R}{S} \right) / \text{Ln}(aN) \quad , \text{ here } a=1$$

When $0 < H < 0.5$, the self-similar correlations at all timescales are anti-persistent, i.e. increases at any time are more likely to be followed by decreases over all later time scales. When $H=0.5$, the self-similar correlations are uncorrelated. When $0.5 < H < 1$, the self-similar correlations at all timescales are persistent, i.e. increases at any time are more likely to be followed by increases over all later time scales.

2) Winter's Exponential Smoothing

In order to take the effects of seasonality and trend into consideration, winter's Exponential Smoothing is used to preliminarily forecast the quantity of sales. For time serial data, Winter's Exponential Smoothing is used to preprocess all the historical data and use them to predict the production demand, which will be entered into the proposed hybrid model as input variable (K4)(see Table 1). Similar to the previous researches, we assume $\alpha = 0.1$, $\beta = 0.1$ and $\gamma = 0.9$. The data generating process is assumed to be of the form

$$x_t = (a_{0,t} + a_{1,t})C_t + \varepsilon_t$$

Where C_t seasonal factor

$$a_{0,t} = \alpha \frac{x_t}{C_{t-N}} + (1 - \alpha)(a_{0,t-1} + a_{1,t-1})$$

is exponentially smoothed level of the process at the end of period t

x_t actual monthly sales in period t

N number of periods in the season ($N=12$ months)

α_{t-1} trend for period $t-1$

α smoothing constant for α_0 .

The season factor, C_t , is updated as follows

$$C_t = \gamma \frac{x_t}{a_{0,t}} + (1 - \gamma)C_{t-N}$$

Where γ is the smoothing constant for C_t . For updating the trend component

$$\alpha_{1,t} = \phi(\alpha_{0,t} - \alpha_{0,t-1}) + (1 - \phi)\alpha_{1,t-1}$$

Where Φ is the smoothing constant for α_1 . Winter's forecasting model is then constructed by

$$\hat{x}_t(\alpha_{0,t} + \alpha_{1,t})C_t$$

Where \hat{x}_t is the estimate in time period t .

C. FCBPN forecasting stage

FCBPN [30] (Fuzzy Clustering and Back-Propagation (BP) Neural Networks with adaptive learning rate) is utilized

to forecast the future packaging industry sales. As shown in figure 2, FCBPN is composed of two steps: (1) utilizing Fuzzy C-Means clustering method (Used in an clusters memberships fuzzy system (CMFS)), the clusters membership levels of each normalized data records will be extracted; (2) Each cluster will be fed into parallel BP networks with a learning rate adapted as the level of cluster membership of training data records.

1) Extract membership levels to each cluster (CMFS)

Using Fuzzy C-Means clustering method (utilized in an adapted fuzzy system (CMFS)), the clusters centers of the normalized data records will be founded, and consequently, we can extract the clusters membership levels of each normalized data records.

1.1) Data normalization

The input values (**K1, K2, K3, K4**) will be ranged in the interval $[0.1, 0.9]$ to meet property of neural networks. The normalized equation is as follows:

$$N_i = 0.1 + 0.8 * (K_i - \min(K_i)) / (\max(K_i) - \min(K_i)). \quad (1)$$

Where K_i presents a key variable, N_i presents normalized input (see Table 1), $\max(K_i)$ and $\min(K_i)$ represent maximum and minimum of the key variables, respectively.

1.2) Fuzzy c-means clustering

In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In Fuzzy c-means (FCM) (developed by Dunn 1973 [14] and improved by Bezdek 1981 [1]), data elements can belong to more than one cluster, and associated with each element is a set of membership levels. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2 \quad , 1 \leq m < \infty.$$

Where u_{ij} is the degree of membership of x_i in the cluster j , x_i is the i^{th} of measured data and c_j is the center of the j^{th} cluster. The algorithm is composed of the following steps:

Step 1: Initialize randomly the degrees of membership matrix $U = [u_{ij}]$, $U^{(0)}$

Step 2: Calculate the centroid for each cluster $C(k) = [c_j]$ with $U(k)$:

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m}$$

Step 3: For each point, update its coefficients of being in the clusters ($U(k)$, $U(k+1)$):

$$u_{ij} = \frac{1}{\sum_{i=1}^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{2/(m-1)}}$$

Step 4: If $\|U^{(k+1)} - U^{(k)}\| < \varepsilon$, $0 < \varepsilon < 1$. then STOP; otherwise return to step 2.

This procedure converges to a local minimum or a saddle point of J_m . According to Bezdek [1], the appreciated parameter combination of two factors (m and ϵ) is $m = 2$ and $\epsilon = 0.5$

Using fuzzy c-means, Table2 shows that the use of four clusters is the best among all different clustering numbers.

Clustering groups	fuzzy c-means total distance
Clustering 2 groups	4.5234
Clustering 3 groups	2.2122
Clustering 4 groups	1.8434

Table 2: Comparison of total distance of different clustering algorithms in.

1.3) The degree of Membership levels (MLC_k)

In this stage, we will use the sigmoid function (Figure3) to improve the precision of results and to accelerate the training process of neural networks. Then, the advanced fuzzy distance between records data (X_i) and a cluster center (C_k) (AFD_k) will be presented as follow:

$$AFD_k(X_i) = \text{sigmoid}(\|C_k - X_i\|, [50, 0.5])$$

$$\text{sigmoid}(x, [a, c]) = \frac{1}{1 + e^{-a(x-c)}}$$

The membership levels of belonging of a record X_i to k^{th} cluster ($MLC_k(X_i)$) is related inversely to the distance from records data X_i to the cluster center C_k ($AFD_k(X_i)$):

$$MLC_k(X_i) = \frac{1}{AFD_k(X_i)}$$

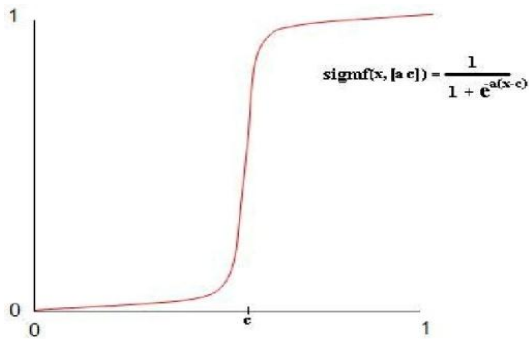


Figure 3: Sigmoid function, a = 50 and c = 0.5.

The clusters memberships' fuzzy system (CMFS) return the memberships level of belonging of data record X to each clusters:

$$CMFS(X) = (MLC_1(X), MLC_2(X), MLC_3(X), MLC_4(X))$$

Thus, we can construct a new training sample ($X_i, MLC_1(X_i), MLC_2(X_i), MLC_3(X_i), MLC_4(X_i)$) for the adaptive neural networks evaluating (Figure2).

2) Adaptive neural networks evaluating stage

The artificial neural networks (ANNs) concept is originated from the biological science (neurons in an organism). Its components are connected according to some pattern of connectivity, associated with different weights. The

weight of a neural connection is updated by learning. The ANNs possess the ability to identify nonlinear patterns by learning from the data set. The back-propagation (BP) training algorithms are probably the most popular ones. The structure of BP neural networks consists of an input layer, a hidden layer, as well as an output layer. Each layer contains $I; J$ and L nodes denoted. The w_{ij} is denoted as numerical weights between input and hidden layers and so is w_{jl} between hidden and output layers as shown in Figure4.

In this stage, we propose an adaptive neural networks evaluating system which consists of four neural networks. Each cluster K is associated with the K^{th} BP network. For each cluster, the training sample will be fed into a parallel Back Propagation networks (BPN) with a learning rate adapted according to the level of clusters membership (MLC_k) of each records of training data set. The structure of the proposed system is shown in Figure 2.

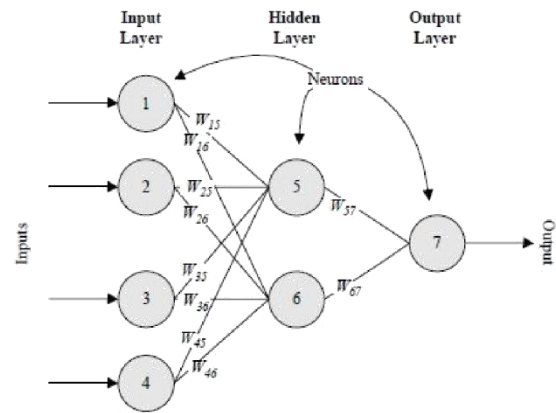


FIGURE 4: The structure of back-propagation neural network

The Adaptive neural networks learning algorithm is composed of two procedures: (a) a feed forward step and (b) a back-propagation weight training step. These two separate procedures will be explained in details as follows:

Step 1- All BP networks are initialized with the same random weights.

Step 2 - Feed forward.

For each BPN_k (associate to the K^{th} cluster), we assume that each input factor in the input layer is denoted by x_i, y_j^k and o_i^k represent the output in the hidden layer and the output layer, respectively. And y_j^k and o_i^k can be expressed as follows:

$$Y_j^k = f(X_j) = f(w_{oj}^k + \sum_{i=0}^I w_{ij}^k x_i)$$

and

$$o_i^k = f(Y_i^k) = f(w_{oi}^k + \sum_{j=1}^J w_{ji}^k y_j^k)$$

where the w_{oj}^k and w_{oi}^k are the bias weights for setting threshold values, f is the activation function used in both hidden and output layers and X_j^k and Y_i^k are the temporarily computing results before applying activation function f . In

this study, a sigmoid function (or logistic function) is selected as the activation function. Therefore, the actual outputs y_j^k and o_l^k in hidden and output layers, respectively, can also be written as:

$$y_j^k = f(x_j^k) = \frac{1}{1 + e^{-x_j^k}}$$

and

$$o_l^k = f(y_l^k) = \frac{1}{1 + e^{-y_l^k}}$$

The activation function f introduces the nonlinear effect to the network and maps the result of computation to a domain (0, 1). In our case, the sigmoid function is used as the activation function.

$$f(t) = \frac{1}{1 + e^{-t}}, f' = f(1 - f)$$

The global output of the adaptive neural networks is calculated as:

$$o_l = \frac{\sum_{k=1}^4 (MLC_k(X_j) \times o_l^k)}{\sum_{k=1}^4 MLC_k(X_j)}$$

As shown above, the effect of the output o_l^k on the global output o_l is both strongly and positively related to the membership level (MLC_k) of data record X_j to k^{th} cluster.

Step 3-Back-propagation weight training. The error function is defined as:

$$E = \frac{1}{2} \sum_{l=1}^L e_l^2 = \frac{1}{2} \sum_{l=1}^L (t_l - o_l)^2$$

Where t_k is a predefined network output (or desired output or target value) and e_k is the error in each output node. The goal is to minimize E so that the weight in each link is accordingly adjusted and the final output can match the desired output. The learning speed can be improved by introducing the momentum term. Usually, falls in the range [0, 1]. For the iteration n and for BPN_k (associated to k^{th} cluster), the adaptive learning rate in BPN_k and the variation of weights Δw_k can be expressed as

$$\eta_k = \frac{MLC_k(X_j)}{\sum_{k=1}^4 MLC_k(X_j)} \times \eta$$

$$\Delta w_k(n+1) = \eta_k \times \Delta w_k(n) + \alpha \times \frac{\delta E}{\delta w_k(n)}$$

As shown above, we can conclude that the variation of the BPN_k network weights (w_{oj}^k and w_{ol}^k) are more important as long as the membership level (MLC_k) of data record X_j to k^{th} cluster is high. If the value of membership level (MLC_k) of data record X_j to k^{th} cluster is close to zero then the changes in BPN_k network weights are very minimal.

The configuration of the proposed BPN is established as follows:

- Number of neurons in the input layer: I = 4
- Number of neurons in the output layer: L = 1
- Single hidden layer

- Number of neurons in the hidden layer: J = 2
- Network-learning rule: delta rule
- Transformation function: sigmoid function
- learning rate: = 0.1
- Momentum constant: = 0.02
- learning times : 20000

IV. EXPERIMENT RESULTS AND ANALYSIS

1) Constructing DELPHI-FCBPN System

The data test used in this study was collected from sales forecasting case study of manufactures packaging In Tangier. The total number of training samples was collected from January 2001 to December 2008 while the total number of testing samples was from January 2009 to December 2009. The proposed DELPHI-FCBPN system was applied as case to forecast the sales. The results are presented in Table 3.

Month	Actual values	Forecasted values
2009/1	5322	5176
2009/2	3796	3766
2009/3	3696	3700
2009/4	5632	5592
2009/5	6126	6097
2009/6	5722	5766
2009/7	6040	6140
2009/8	6084	6084
2009/9	7188	7088
2009/10	7079	7079
2009/11	7287	7376
2009/12	7461	7532

Table 3: The forecasted results by DELPHI-FCBPN method.

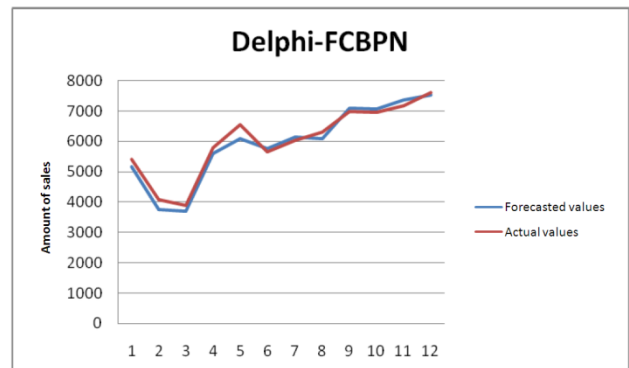


Figure 5: The MAPE of DELPHI-FCBPN.

4.2. Comparisons of FCBPN model with other previous models

Experimental comparison of outputs of DELPHI-FCBPN with other methods shows that the proposed model outperforms the previous approaches (Tables 4-6). We apply two different performance measures called mean absolute percentage error (MAPE) and root mean square error (RMSE), to compare the FCBPN model with the previous methods: BPN, WES and FNN.

Month	Actual values	BPN forecasts
2009/1	5408	5437
2009/2	4089	3512
2009/3	3889	3880
2009/4	5782	5865
2009/5	6548	6064
2009/6	5660	6403
2009/7	6032	6181
2009/8	6312	6329
2009/9	6973	7159
2009/10	6941	7427
2009/11	7174	7342
2009/12	7601	8100

Table4: The forecasted results by BPN method

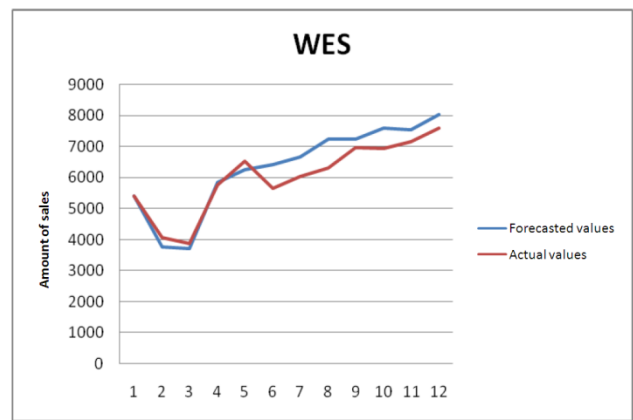


Figure 7: The MAPE of WES

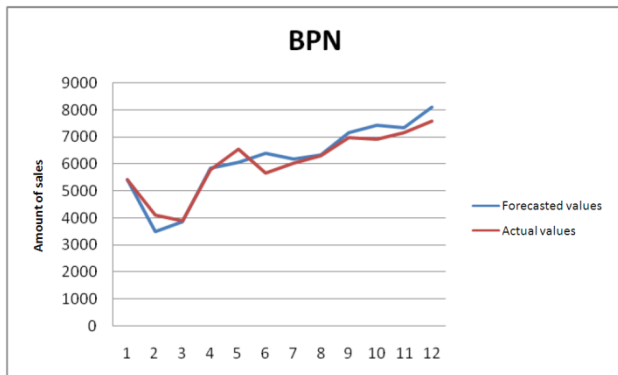


Figure 6: The MAPE of BPN.

Month	Actual values	Forecasted values
2009/1	5408	5399
2009/2	4089	3774
2009/3	3889	3722
2009/4	5782	5836
2009/5	6548	6252
2009/6	5660	6412
2009/7	6032	6658
2009/8	6312	7244
2009/9	6973	7233
2009/10	6941	7599
2009/11	7174	7543
2009/12	7601	8022

Table5: The forecasted results by Winter's method.

Month	Actual values	Forecasted values
2009/1	5408	5131
2009/2	4089	3562
2009/3	3889	3589
2009/4	5782	5651
2009/5	6548	6504
2009/6	5660	5905
2009/7	6032	6066
2009/8	6312	6123
2009/9	6973	6968
2009/10	6941	7489
2009/11	7174	7365
2009/12	7601	7768

Table6: The forecasted results by FNN method.

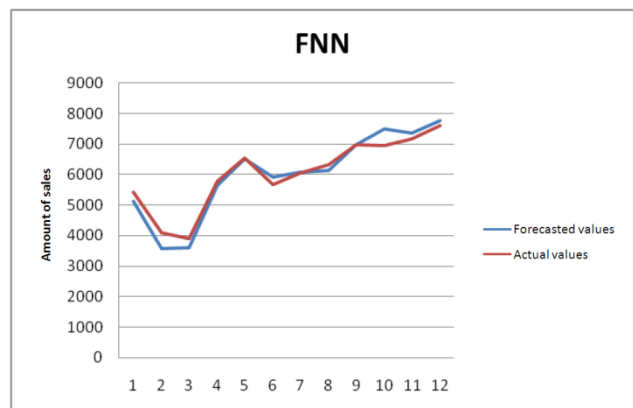


Figure 8: The MAPE of FNN.

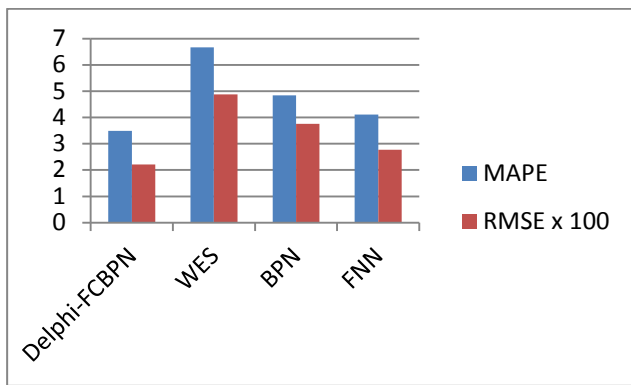


Figure 9: Summary MAPE and RMSE values of prediction methods Delphi-FCBPN, WES, BPN and FNN

Methods	precision	
	MAPE	RMSE
Delphi-FCBPN	3,49	221
WES	6,66	488
BPN	4,85	376
FNN	4,11	278

Table7: Summary MAPE and RMSE values of prediction methods Delphi-FCBPN, WES, BPN and FNN.

$$MAPE = 100 \times \frac{1}{N} \sum_{t=1}^N \frac{|Y_t - P_t|}{Y_t}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (Y_t - P_t)^2}$$

Where, P_t is the expected value for period t , Y_t is the actual value for period t and N is the number of periods.

As shown in Table7, the forecasting accuracy of Delphi-FCBPN is superior to the other traditional approaches regarding MAPE and RMSE evaluations which are summarized in Table7.

V. CONCLUSION

Recently, more and more researchers and industrial practitioners are interested in applying fuzzy theory and neural network in their routine problem solving. This research combines fuzzy theory and back-propagation network into a hybrid system, which will be applied in the sales forecasting of packaging industries sales. This research proposes a hybrid system based on Delphi method, fuzzy clustering and Back-propagation Neural Networks with adaptive learning rate (FCBPN) for sales forecasting.

We applied DELPHI-FCBPN for sales forecasting in a case study of manufactures packaging in Tangier. The experimental results of the proposed approach in Section 4 demonstrated that the effectiveness of the DELPHI-FCBPN is superior to the previous and traditional approaches: WES, BPN and FNN regarding MAPE and RMSE evaluations.

REFERENCES

[1] Bezdek J. C. (1981) : "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York.

[2] Casillas, J., Cordón, O., Herrera, F., & Villar, P. (2004). A hybrid learning process for the knowledge base of a fuzzy rule-based system. In Proceedings of the 2004 international conference on information processing and management of uncertainty in knowledge-based systems, Perugia, Italy (pp. 2189-2196).

[3] Chang, P.-C., & Lai, C.-Y. (2005). A hybrid system combining self-organizing maps with case-based reasoning in wholesaler's new-release book forecasting. *Expert Systems with Applications*, 29(1), 183-192.

[4] Chang, P.-C., & Liu, C.-H. (2008). A TSK type fuzzy rule based system for stock price prediction. *Expert Systems with Applications*, 34, 135-144.

[5] Chang, P.-C., Liu, C.-H., & Fan, C. Y. (2009). Data clustering and fuzzy neural network for sales forecasting: A case study in printed circuit board industry. *Knowledge Based Systems*, 22(5), 344-355.

[6] Chang, P.-C., Liu, C.-H., & Wang, Y.-W. (2006). A hybrid model by clustering and evolving fuzzy rules for sales decision supports in printed circuit board industry. *Decision Support Systems*, 42(3), 1254-1269.

[7] Chang, P.-C., & Wang, Y.-W. (2006). Fuzzy Delphi and back-propagation model for sales forecasting in PCB industry. *Expert Systems with Applications*, 30(4).

[8] Chang, P.-C., Wang, Y.-W., & Liu, C.-H. (2007). The development of a weighted evolving fuzzy neural network for PCB sales forecasting. *Expert Systems with Applications*, 32(1), 86-96.

[9] Chang, P.-C., Wang, Y.-W., & Tsai, C.-Y. (2005). Evolving neural network for printed circuit board sales forecasting. *Expert Systems with Applications*, 29, 83-92.

[10] Chang, P.-C., Yen-Wen Wang, Chen-Hao Liu (2007). The development of a weighted evolving fuzzy neural network for PCB sales forecasting. *Expert Systems with Applications*, 32(1), 86-96.

[11] Chang Pei-Chann, Wang Di-di, Zhou Changle (2011). A novel model by evolving partially connected neural network for stock price trend forecasting. *Expert Systems with Applications*, Volume 39 Issue 1, January, 2012.

[12] Cordón, O., & Herrera, F. (1997). A three-stage evolutionary process for learning descriptive and approximate fuzzy logic controller knowledge bases from examples. *International Journal of Approximate Reasoning*, 17(4), 369-407.

[13] Cordón, O., Herrera, F., Hoffmann, F., & Magdalena, L. (2001). Genetic fuzzy systems: Evolutionary tuning and learning of fuzzy knowledge bases. Singapore: World Scientific.

[14] Dunn J. C. (1973) : "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", *Journal of Cybernetics* 3 : 32-57.

[15] Eberhart, R. C., and Kennedy, J. (1995). A new optimizer using particle swarm theory. *Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, Nagoya, Japan, 39-43. Piscataway, NJ : IEEE Service Center.

[16] Efendigil, T., Önu, S., & Kahraman, C. (2009). A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis. *Expert Systems with Applications*, 36(3), 6697-6707.

[17] Esmir, A. (2007). Generating Fuzzy Rules from Examples Using the Particle Swarm Optimization Algorithm. *Hybrid Intelligent Systems*, 2007. HIS 2007. 7th International Conference IEEE.

[18] Hadavandi Esmaeil, Shavandi Hassan, & Ghanbari Arash (2011). An improved sales forecasting approach by the integration of genetic fuzzy systems and data clustering : Case study of printed circuit board. *Expert Systems with Applications* 38 (2011) 9392-9399.

[19] Goldberg, D. A. (1989). Genetic algorithms in search, optimization, and machine learning. Reading, MA: Addison-Wesley.

[20] Kuo, R. J., & Chen, J. A. (2004). A decision support system for order selection in electronic commerce based on fuzzy neural network supported by real-coded genetic algorithm. *Expert Systems with Applications*, 26(2), 141-154.

[21] Kuo, R.J. & Xue, K.C. (1999). Fuzzy neural networks with application to sales forecasting. *Fuzzy Sets and Systems*, 108, 123-143.

[22] KyongJoo Oh, Ingoo Han (2001). An intelligent clustering forecasting system based on change-point detection and artificial neural networks

- :application to financial economics. System Sciences, 2001. Proceedings of the 34th Annual Hawaii
- [23] Martínez-López, F., & Casillas, J. (2009). Marketing intelligent systems for consumer behavior modelling by a descriptive induction approach based on genetic fuzzy. *Industrial Marketing Management*.
- [24] Orriols-Puig, A., Casillas, J., & Martínez-López, F. (2009). Unsupervised learning of fuzzy association rules for consumer behavior modeling. *Mathware and Soft Computing*, 16(1), 29-43.
- [25] SaeidIranmanesh, M. Amin Mahdavi (2009). A Differential Adaptive Learning Rate Method for Back-Propagation Neural Networks. *World Academy of Science, Engineering and Technology* 50 2009.
- [26] Sivanandam.S. N, VisalakshiP : Dynamic task scheduling with load balancing using parallel orthogonal particle swarm optimisation. *IJBIC* 1(4) : 276-286 (2009)
- [27] Toly Chen (2003). A fuzzy back propagation network for output time prediction in a wafer fab. *Applied Soft Computing* 2/3F (2003), 211-222.
- [28] Zhang, Wang & Chang (2009) . A Model on Forecasting Safety Stock of ERP Based on BP Neural Network. *Proceedings of the 2008 IEEE ICMIT* [29] Zhang, Haifeng and Huang(2010), Sales Forecasting Based on ERP System through BP Neural Networks. *ISICA'10 Proceedings of the 5th international conference on Advances in computation and intelligence*
- [29] Chang, I-C., Hwang, H-G., Hung, M-C., Chen, S-L., Yen, D., "A Neural Network Evaluation Model for ERP Performance from SCM Perspective to Enhance Enterprise Competitive Advantage, [J]." *Expert Systems with Applications* 2007.08:102
- [30] Attariuas, Bouhorma and el Fallahi "An improved approach based on fuzzy clustering and Back-Propagation Neural Networks with adaptive learning rate for sales forecasting: Case study of PCB industry". *IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 3, No 1, May 2012 ISSN (Online): 1694-0814

AUTHORS PROFILE

AttariuasHicham received the computer engineer degree in 2009 from **ENSAIS** national school of computer science and systems analysis in Rabat, Morocco. Currently, he is a PhD Student in Computer Science. Current research interests: fuzzy system, intelligence system, bac-propagation network, genetic intelligent system.

Bouhorma Mohammed received the PhD degree in Telecommunications and Computer Engineering. He is a Professor of Telecommunications and Computer Engineering in Abdelmalek Essaadi University. He has been a member of the Organizing and the Scientific Committees of several symposia and conferences dealing with intelligent system, Mobile Networks, Telecommunications technologies.

SofiAnas received the Master degree in Telecommunications from AbdelmalekEssaadi University. He is a PhD Student in Telecommunications and intelligent system. Current research interests: Mobile Networks, Telecommunications technologies and intelligent system.

Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches

Md. Mijanur Rahman

Dept. of Computer Science & Engineering,
Jatiya Kabi Kazi Nazrul Islam University
Trishal, Mymensingh, Bangladesh.

Md. Al-Amin Bhuiyan

Dept. of Computer Science & Engineering,
Jahangir nagar University
Savar, Dhaka, Bangladesh.

Abstract—This paper presents simple and novel feature extraction approaches for segmenting continuous Bangla speech sentences into words/sub-words. These methods are based on two simple speech features, namely the time-domain features and the frequency-domain features. The time-domain features, such as short-time signal energy, short-time average zero crossing rate and the frequency-domain features, such as spectral centroid and spectral flux features are extracted in this research work. After the feature sequences are extracted, a simple dynamic thresholding criterion is applied in order to detect the word boundaries and label the entire speech sentence into a sequence of words/sub-words. All the algorithms used in this research are implemented in Matlab and the implemented automatic speech segmentation system achieved segmentation accuracy of 96%.

Keywords—Speech Segmentation; Features Extraction; Short-time Energy; Spectral Centroid; Dynamic Thresholding.

I. INTRODUCTION

Automated segmentation of speech signals has been under research for over 30 years [1]. Speech Recognition system requires segmentation of Speech waveform into fundamental acoustic units [2]. Segmentation is a process of decomposing the speech signal into smaller units. Segmentation is the very basic step in any voiced activated systems like speech recognition system and speech synthesis system. Speech segmentation was done using wavelet [3], fuzzy methods [4], artificial neural networks [5] and Hidden Markov Model [6]. But it was found that results still do not meet expectations. In order to have results more accurate, groups of several features were used [7, 8, 9 and 10]. This paper is continuation of feature extraction for speech segmentation research. The method implemented here is a very simple example of how the detection of speech segments can be achieved.

This paper is organized as follows: Section 2 describes techniques for segmentation of the speech signal. In Section 3, we will describe different short-term speech features. In section 4, the methodological steps of the proposed system will be discussed. Section 5 and 6 will describe the experimental results and conclusion, respectively.

II. SPEECH SEGMENTATION

Automatic speech segmentation is a necessary step that used in Speech Recognition and Synthesis systems. Speech segmentation is breaking continuous streams of sound into some basic units like words, phonemes or syllables that can be recognized. The general idea of segmentation can be described

as dividing something continuous into discrete, non-overlapping entities [11]. Segmentation can be also used to distinguish different types of audio signals from large amounts of audio data, often referred to as *audio classification* [12].

Automatic speech segmentation methods can be classified in many ways, but one very common classification is the division to *blind* and *aided segmentation algorithms*. A central difference between aided and blind methods is in how much the segmentation algorithm uses previously obtained data or external knowledge to process the expected speech. We will discuss about these two approaches in the following sub-sections.

A. Blind segmentation

The term *blind* segmentation refers to methods where there is no pre-existing or external knowledge regarding linguistic properties, such as orthography or the full phonetic annotation, of the signal to be segmented. Blind segmentation is applied in different applications, such as speaker verification systems, speech recognition systems, language identification systems, and speech corpus segmentation and labeling [13].

Due to the lack of external or top-down information, the first phase of blind segmentation relies entirely on the acoustical features present in the signal. The second phase or bottom-up processing is usually built on a front-end parametrization of the speech signal, often using MFCC, LP-coefficients, or pure FFT spectrum [14].

B. Aided segmentation

Aided segmentation algorithms use some sort of external linguistic knowledge of the speech stream to segment it into corresponding segments of the desired type. An orthographic or phonetic transcription is used as a parallel input with the speech, or training the algorithm with such data [15]. One of the most common methods in ASR for utilizing phonetic annotations is with HMM-based systems [16]. HMM-based algorithms have dominated most speech recognition applications since the 1980's due to their so far superior performance in recognition and relatively small computational complexity in the field of speech recognition [17].

III. FEATURE EXTRACTION FOR SPEECH SEGMENTATION

The segmentation method described here is a purely bottom-up blind speech segmentation algorithm. The general principle of the algorithm is to track the amplitude or spectral changes in the signal by using short-time energy or spectral

features and to detect the segment boundaries at the locations where amplitude or spectral changes exceed a minimum threshold level. Two types of features are used for segmenting speech signal: *time-domain signal features* and *frequency-domain signal features*.

A. Time-Domain Signal Features

Time-domain features are widely used for speech segment extraction. These features are useful when it is needed to have algorithm with simple implementation and efficient calculation. The most used features are short-time energy and short-term average zero-crossing rate.

1) Short-Time Signal Energy

Short-term energy is the principal and most natural feature that has been used. Physically, energy is a measure of how much signal there is at any one time. Energy is used to discover voiced sounds, which have higher energy than silence/un-voiced, in a continuous speech, as shown in Figure-1.

The energy of a signal is typically calculated on a short-time basis, by windowing the signal at a particular time, squaring the samples and taking the average [18]. The square root of this result is the engineering quantity, known as the root-mean square (RMS) value, also used. The short-time energy function of a speech frame with length N is defined as

$$E_n = \frac{1}{N} \sum_{m=1}^N [x(m)w(n-m)]^2 \dots\dots\dots (1)$$

The short-term root mean squared (RMS) energy of this frame is given by:

$$E_{n(RMS)} = \sqrt{\frac{1}{N} \sum_{m=1}^N [x(m)w(n-m)]^2} \dots\dots\dots (2)$$

Where $x(m)$ is the discrete-time audio signal and $w(m)$ is rectangle window function, given by the following equation:

$$w(m) = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & \text{Otherwise} \end{cases} \dots\dots\dots (3)$$

2) Short-Time Average Zero-Crossing Rate

The average zero-crossing rate refers to the number of times speech samples change algebraic sign in a given frame. The rate at which zero crossings occur is a simple measure of the frequency content of a signal. It is a measure of number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero [19]. Unvoiced speech components normally have much higher ZCR values than voiced ones, as shown in Figure 2. The short-time average zero-crossing rate is defined as

$$Z_n = \frac{1}{2} \sum_{m=1}^N |\text{sgn}[x(m)] - \text{sgn}[x(m-1)]| w(n-m) \dots (4)$$

Where, $\text{sgn}[x(m)] = \begin{cases} 1 & x(m) \geq 0 \\ -1 & x(m) < 0 \end{cases} \dots\dots\dots (5)$

and $w(n)$ is a rectangle window of length N , given in equation (3).

B. Frequency-Domain Signal Features

The most information of speech is concentrated in 250Hz – 6800Hz frequency range [20]. In order to extract frequency-domain features, discrete Fourier transform (that provides information about how much of each frequency is present in a signal) can be used. The Fourier representation of a signal shows the spectral composition of the signal. Widely used frequency-domain features are *spectral centroid* and *spectral flux* feature sequences that used discrete Fourier transform.

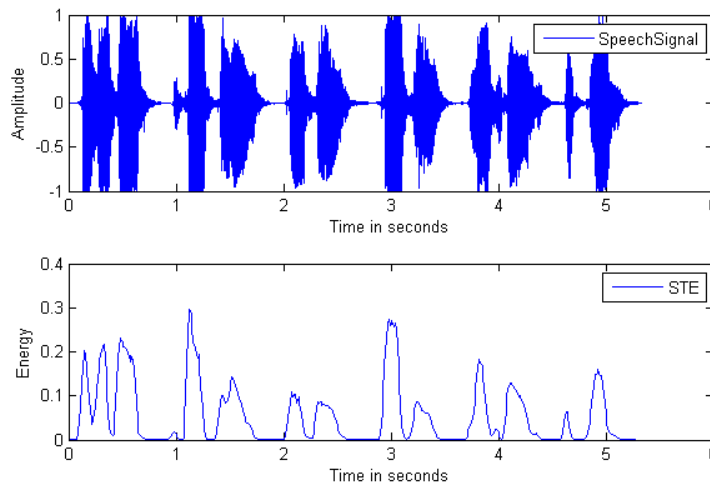


Figure 1. Original signal and short-time energy curves of the speech sentence, “আমাদেরজাতীয়কবিকাজীনজরুলহসলাম”.

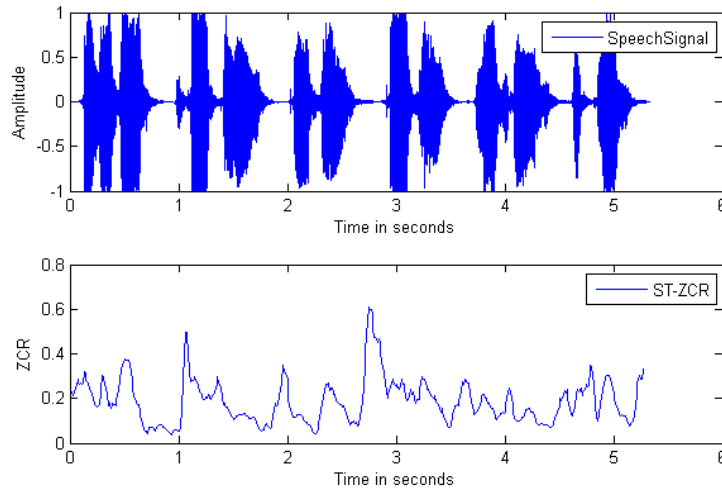


Figure 2. Original signal and short-term average zero-crossing rate curves of speech sentence, “আমাদেরজাতীয়কবিকাজীনজরুলইসলাম”.

1) Spectral Centroid

The spectral centroid is a measure used in digital signal processing to characterize a spectrum. It indicates where the "center of gravity" of the spectrum is. This feature is a measure of the spectral position, with high values corresponding to “brighter” sounds [21], as shown in Figure-3. The spectral centroid, SC_i , of the i -th frame is defined as the center of “gravity” of its spectrum and it is given by the following equation:

$$SC_i = \frac{\sum_{m=0}^{N-1} f(m)X_i(m)}{\sum_{m=0}^{N-1} X_i(m)} \dots\dots\dots (6)$$

Here, $f(m)$ represents the center frequency of i -th bin with length N and $X_i(m)$ is the amplitude corresponding to that bin in DFT spectrum. The DFT is given by the following equation and can be computed efficiently using a fast Fourier transform (FFT) algorithm [22].

$$X_k = \sum_{n=0}^{N-1} x(n)e^{-j2\pi k \frac{n}{N}} ; k=0, \dots, N-1 \dots\dots\dots (7)$$

2) Spectral flux

Spectral flux is a measure of how quickly the power spectrum of a signal is changing (as shown in Figure 4), calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame, also known as the Euclidean distance between the two normalized spectra. The spectral flux can be used to determine the timbre of an audio signal, or in onset detection [23], among other things. The equation of Spectral Flux, SF_i is given by:

$$SF_i = \sum_{k=1}^{N/2} (|X_i(k)| - |X_i(k-1)|)^2 \dots\dots\dots (8)$$

Here, $X_i(k)$ is the DFT coefficients of i -th short-term frame with length N , given in equation (7).

C. Speech Segments Detection

After computing speech feature sequences, a simple dynamic threshold-based algorithm is applied in order to detect the speech word segments. The following steps are included in this thresholding algorithm.

1. Get the feature sequences from the previous feature extraction module.
2. Apply median filtering in the feature sequences.
3. Compute the Mean or average values of smoothed feature sequences.
4. Compute the histogram of the smoothed feature sequences.
5. Find the local maxima of histogram.
6. If at least two maxima M_1 and M_2 have been found, Then:

$$\text{Threshold, } T = \frac{W * M_1 + M_2}{W + 1} \dots\dots\dots (9)$$

Otherwise,

$$\text{Threshold, } T = \frac{\text{Mean}}{2} \dots\dots\dots (10)$$

where W is a user-defined weight parameter [24]. Large values of W obviously lead to threshold values closer to M_1 . Here, we used $W=100$.

The above process is applied for both feature sequences and finding two thresholds: T_1 based on the energy sequences and T_2 on the spectral centroid sequences. After computing two thresholds, the speech word segments are formed by successive frames for which the respective feature values are larger than the computed thresholds (for both feature sequences). Figure-5 shows both filtered feature sequences curves with threshold values.

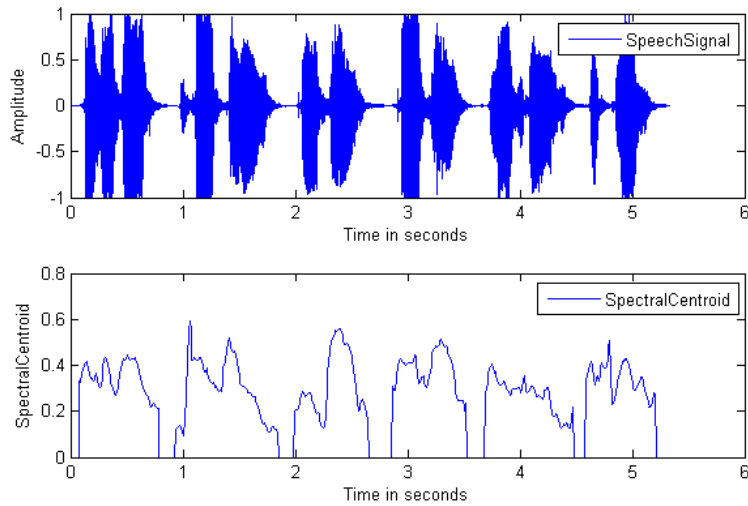


Figure 3. The graph of original signal and spectral centroid features of speech sentence, “আমাদেরজাতীয়কবিকাজীনজরুলইসলাম”.

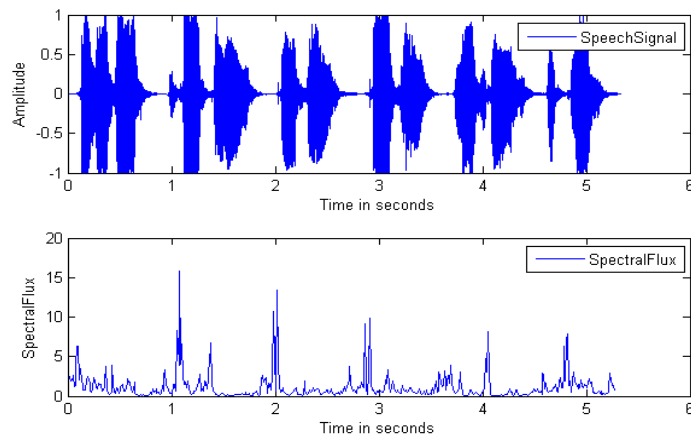


Figure 4. The curves of original signal and spectral flux features of speech sentence, “আমাদেরজাতীয়কবিকাজীনজরুলইসলাম”.

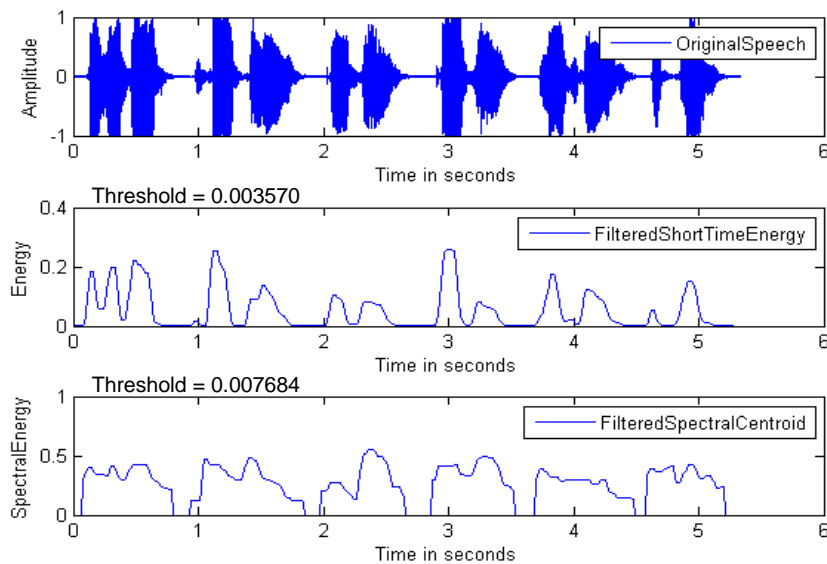


Figure 5. Original speech signal and median filtered feature sequences curves with threshold values of a speech sentence “আমাদেরজাতীয়কবিকাজীনজরুলইসলাম”.

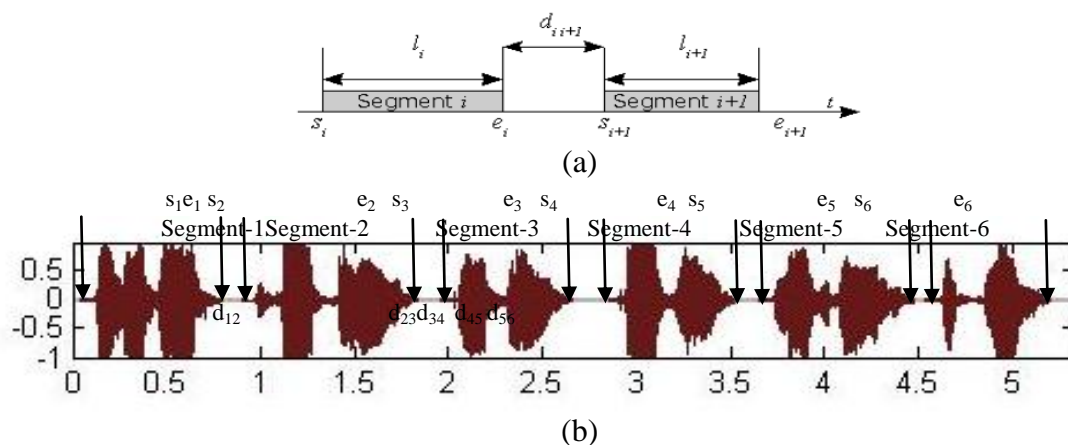


Figure 6. Speech Segments:(a) Segment, where l_i is the length of i segment, d_{i+1} is the distance between two segments and s_i is start time of i segment, e_i is the end time of i segment. (b) Detection of start and end point of each segment in a speech sentence.

D. Post Processing of Detected Segments

As shown in Figure 6, detected speech segments are analyzed in post-processing stage. Common segmentation errors are: short segments usually are noises/silences, and two segments with short space in between can be the same segment.

Post-processing with rule base can fix these and similar mistakes. Waheed [25] proposed to use 2 rules:

1. If $l_i < \min Length$ and $d_{i+1} > \min Space$, then the segment i is discarded, similarly if $l_{i+1} < \min Length$ and $d_{i+1} > \min Space$, then segment $i+1$ is discarded.
2. If l_i or $l_{i+1} > \min Length$ and $d_{i+1} > \min Space$ and $l_i + l_{i+1} > FL$, then two segments are merged and anything between the two segments that was previously left, is made part of the speech.

IV. IMPLEMENTATION

The automatic speech segmentation system has been implemented in Windows environment and we have used MATLAB Tool Kit for developing this application. The proposed speech segmentation system has six major steps, as shown in Figure 7.

- A. Speech Acquisition
- B. Signal Preprocessing
- C. Speech Features Extraction
- D. Histogram Computation
- E. Dynamic thresholding and
- F. Post-Processing

A. Speech Acquisition

Speech acquisition is acquiring of continuous Bangla speech sentences through the microphone.

Speech capturing or speech recording is the first step of implementation. Recording has been done by native male speaker of Bangali. The sampling frequency is 16 KHz; sample size is 8 bits, and mono channels are used.

B. Signal Preprocessing

This step includes elimination of background noise, framing and windowing. Background noise is removed from the data so that only speech samples are the input to the further processing. Continuous speech signal has been separated into a number of segments called frames, also known as framing. After the pre-emphasis, filtered samples have been converted into frames, having frame size of 50 msec. Each frame overlaps by 10 msec. To reduce the edge effect of each frame segment windowing is done. The window, $w(n)$, determines the portion of the speech signal that is to be processed by zeroing out the signal outside the region of interest. Rectangular window has been used.

C. Short-term Feature Extraction

After windowing, we have been computed the short-term energy features and spectral centroid features of each frame of the speech signal. These features have been discussed in details in Section 3. In this step, median filtering of these feature sequences also computed.

D. Histogram Computation

Histograms of both smoothed feature sequences are computed in order to find the local maxima of the histogram, from which the threshold values are calculated.

E. Dynamic Thresholding

Dynamic thresholding is applied for both feature sequences and finding two thresholds: T_1 and T_2 , based on the energy sequences and the spectral centroid sequences respectively.

After computing two thresholds, the speech word segments are formed by successive frames for which the respective feature values are larger than the computed thresholds.

F. Post-Processing

In order to segment words/sub-words, the detected speech segments are lengthened by 5 short term windows (each

window of 50 msec), on both sides in the post-processing step. Two segments with short space in between have been merged to get final speech segments. These segmented speech words are saved as *.wav file format for further use.

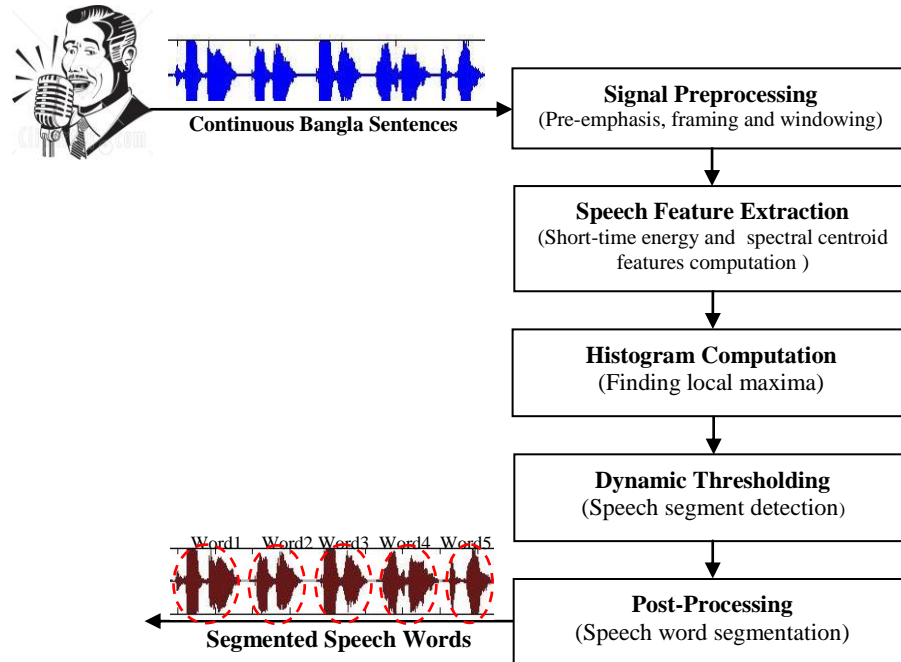


Figure 7. Block diagram of the proposed Automatic Speech Segmentation system.

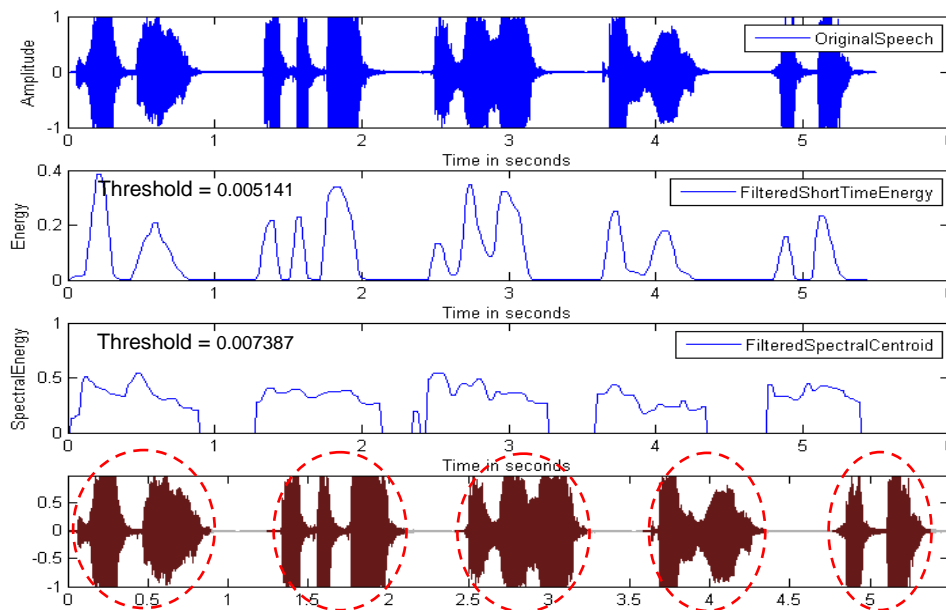


Figure-8. The segmentation results for a speech sentences “জাতীয়পতাকাডিজাইনারকামরুলহাসান” which contains 5 (five) speech words. The first subfigure shows the original signal. The second subfigure shows the sequences of the signal’s energy. In the third subfigure the spectral centroid sequence is presented. In both cases, the respective thresholds are also shown. The final subfigure presents the segmented words in dashed circles.

TABLE 1. SEGMENTATION RESULTS

Sentence ID	No. of word segments expected	No. of words properly segmented by system	Segmentation rate (%)	
			Success rate	Failure rate
S1	6	6	100	0
S2	10	10	100	0
S3	9	9	100	0
S4	5	4	80	20
S5	10	10	100	0
S6	7	7	100	0
S7	8	8	100	0
S8	11	10	90.9	9.1
S9	9	8	88.89	11.11
S10	5	5	100	0
Total	80	77	96.25	3.75

V. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed system different experiments were carried out. All the techniques and algorithms discussed in this paper have been implemented in Matlab 7.12.0 version. In this experiment, various speech sentences in Bangla language have been recorded, analyzed and segmented by using time-domain and frequency-domain features with dynamic thresholding technique. Figure 8 shows the filtered short-time energy and spectral centroid features of the Bangla speech sentence “জাতীয়পতাকারডিজিহনারকামরুলহাসান”, where the boundaries of words are marked automatically by the system. Table-1 shows the details segmentation results for ten speech sentences and reveals that the average segmentation accuracy rate is 96.25%, and it is quite satisfactory.

VI. CONCLUSION AND FURTHER RESEARCH

We have presented a simple speech features extraction approach for segmenting continuous speech into word/sub-words in a simple and efficient way. The short-term speech features have been selected for several reasons. *First*, it provides a basis for distinguishing voiced speech components from unvoiced speech components, i.e., if the level of background noise is not very high, the energy of the voiced segments is larger than the energy of the silent or unvoiced segments. *Second*, if unvoiced segments simply contain environmental sounds, then the spectral centroid for the voiced segments is again larger. *Third*, its change pattern over the time may reveal the rhythm and periodicity nature of the underlying sound.

From the experiments, it was observed that some of the words were not segmented properly. This is due to different causes: (i) the utterance of words and sub-words differs depending on their position in the sentence, (ii) the pauses between the words or sub-words are not identical in all cases because of the variability of the speech signals and (ii) the non-uniform articulation of speech. Also, the speech signal is very much sensitive to the speaker's properties such as age, sex, and emotion. The proposed approach shows good results

in speech segmentation that achieves about 96% of segmentation accuracy. This reduces the memory requirement and computational time in any speech recognition system.

The major goal of future research is to search for possible mechanisms that can be employed to enable top-down feedback and ultimately pattern discovery by learning. To design more reliable system, future systems should employ knowledge (syntactic or semantic) of linguistics and more powerful recognition approaches like Gaussian Mixture Models (GMMs), Time-Delay Neural Networks (TDNNs), Hidden Markov Model (HMM), Fuzzy logic, and so on.

REFERENCES

- [1] OkkoRasanen, “Speech Segmentation and Clustering Methods for a New Speech Recognition Architecture”, M.Sc Thesis, Department of Electrical and Communications Engineering, Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Espoo, November 2007.
- [2] R. Thangarajan, A. M. Natarajan, M. Selvam, “Syllable modeling in continuous speech recognition for Tamil language”, International Journal of Speech Technology, vol. 12, no. 1, pp. 47-57, 2009.
- [3] Hioka Y and Namada N, “Voice activity detection with array signal processing in the wavelet domain”, IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, 86(11):2802-2811, 2003.
- [4] Beritelli F and Casale S, “Robust voiced/unvoiced classification using fuzzy rules”, In 1997 IEEE workshop on speech coding for telecommunications proceeding, pages5-6, 1997.
- [5] Qi Y and Hunt B, “Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier”, IEEE Transactions on Speech and Audio Processing, I(2):250-255, 1993.
- [6] Basu S, “A linked-HMM model for robust voicing and speech detection”, In IEEE international conference on acoustics, speech and signal processing (ICAASSP'03), 2003.
- [7] Atal B and Rabiner L, “A pattern recognition approach to voice-unvoiced-silence classification with applications to speech recognition”, IEEEASSP, ASSP-24(3):201-212, June 1976.
- [8] Siegel L and Bessey A, “Voiced/unvoiced/mixed excitation classification of speech”, IEEE Transactions on Acoustics, Speech and Signal Processing, 30(3):451-460, 1982.
- [9] Shin W, Lee B, Lee Y and Lee J, “Speech/Non-speech classification using multiple features for robust endpoint detection”, In 2000 IEEE International Conference on Acoustics, Speech and Signal Processing,

- ICASSP'00 Proceedings, Vol.3, 2000.
- [10] Kida Y and Kawahara T, "Voice activity detection based on optimally weighted combination of multiple features", In 9th European Conference on Speech Communication and Technology, ISCA, 2005.
- [11] Kvale K, "Segmentation and Labeling of Speech", PhD Dissertation, The Norwegian Institute of Technology, 1993.
- [12] Antal M, "Speaker Independent Phoneme Classification in Continuous Speech", Studia Univ. Babeş-Bolyai, Informatica, Vol. 49, No. 2, 2004.
- [13] Sharma M and Mammon R, "Blind speech segmentation: Automatic segmentation of speech without linguistic knowledge", Spoken Language, 1996. ICSLP 96. Proceedings. Vol. 2, pp. 1237-1240, 1996.
- [14] Sai Jayram A K V, Ramasubramanian V and Sreenivas T V, "Robust parameters for automatic segmentation of speech", Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '02), Vol. 1, pp. 513-516, 2002.
- [15] Schiel F, "Automatic Phonetic Transcription of Non-Prompted Speech", Proceedings of the ICPhS 1999. San Francisco, August 1999. pp. 607-610, 1999.
- [16] Knill K and Young S, "Hidden Markov Models in Speech and Language Processing", Kluwer Academic Publishers, pp. 27-68, 1997.
- [17] Juang B H and Rabiner L R, "Automatic Speech Recognition – A Brief History of The Technology Development", Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2005.
- [18] Tong Zhang and Jay C Kuo, "Hierarchical classification of audio data for archiving and retrieving", In International Conference on Acoustics, Speech and Signal Processing, volume VI, pages 3001–3004. IEEE, 1999.
- [19] L R Rabiner and M R Sambur, "An Algorithm for determining the endpoints of Isolated Utterances", The Bell System Technical Journal, February 1975, pp 298-315.
- [20] Niederjohn R and Grotelueschen J, "The enhancement of speech intelligibility in high noise level by high-pass filtering followed by rapid amplitude compression", IEEE Transactions on Acoustics, Speech and Signal Processing, 24(4), pp277-282, 1976.
- [21] T Giannakopoulos, "Study and application of acoustic information for the detection of harmful content and fusion with visual information" Ph.D. dissertation, Dept. of Informatics and Telecommunications, University of Athens, Greece, 2009.
- [22] Cooley, James W. and Tukey, John W. "An algorithm for the machine calculation of complex Fourier series", Mathematics of Computation: Journal Review, 19: 297–301, 1965.
- [23] Bello J P, Daudet L, Abdallah S, Duxbury C, Davies M, and Sandler MB, "A Tutorial on Onset Detection in Music Signals", IEEE Transactions on Speech and Audio Processing 13(5), pp 1035–1047, 2005.
- [24] T Giannakopoulos, A Pikrakis and S. Theodoridis "A Novel Efficient Approach for Audio Segmentation", Proceedings of the 19th International Conference on Pattern Recognition (ICPR2008), December 8-11 2008, Tampa, Florida, USA.
- [25] Waheed K, Weaver K and Salam F, "A robust algorithm for detecting speech segments using an entropic contrast", In Proc. of the IEEE Midwest Symposium on Circuits and Systems, Vol.45, Lida Ray Technologies Inc., 2002.

AUTHORS PROFILE

Md. MijanurRahman



Mr. Md. Mijanur Rahman is working as Assistant Professor of the Dept. of Computer Science and Engineering at Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymen singh, Bangladesh. Mr. Rahman obtained his B. Sc. (Hons) and M. Sc degrees, both with first class first in Computer Science and Engineering from Islamic University, Kushtia, Bangladesh. Now he is a PhD researcher of the department of Computer Science and Engineering at Jahangir

nagar University, Savar, Dhaka, Bangladesh. His teaching and research interest lies in the areas such as Digital Speech Processing, Pattern Recognition, Database management System, Artificial Intelligence, etc. He has got many research articles published in both national and international journals.

Dr. Md. Al-Amin Bhuiyan



Dr. Md. Al-Amin Bhuiyan is serving as a professor of the Dept. of Computer Science and Engineering. Dr. Bhuiyan obtained his B. Sc. (Hons) and M. Sc degrees both with first class in Applied Physics & Electronics from University of Dhaka, Bangladesh. He completed his PhD study in Information & Communication Engineering from Osaka City University, Japan.

His teaching and research interest lies in the areas such as Image Processing, Computer Vision, Computer Graphics, Pattern Recognition, Soft Computing, Artificial Intelligence, Robotics, etc. He has got many articles published in both national and international journals.

A Randomized Fully Polynomial-time Approximation Scheme for Weighted Perfect Matching in the Plane

Yasser M. Abd El-Latif
Faculty of Science,
Ain Shams University,
Cairo, Egypt

Salwa M. Ali
Faculty of Science,
Ain Shams University,
Cairo, Egypt

Hanaa A.E. Essa
Faculty of Science,
Tanta University,
Tanta, Egypt

Soheir M. Khamis
Faculty of Science,
Ain Shams University,
Cairo, Egypt

Abstract— In the approximate Euclidean min-weighted perfect matching problem, a set V of $2n$ points in the plane and a real number $\varepsilon > 0$ are given. Usually, a solution of this problem is a partition of points of V into n pairs such that the sum of the distances between the paired points is at most $(1 + \varepsilon)$ times the optimal solution.

In this paper, the authors give a randomized algorithm which follows a Monte-Carlo method. This algorithm is a randomized fully polynomial-time approximation scheme for the given problem. Fortunately, the suggested algorithm is a one tackled the matching problem in both Euclidean nonbipartite and bipartite cases.

The presented algorithm outlines as follows: With repeating $\lceil 1/\varepsilon \rceil$ times, we choose a point from V to build the suitable pair satisfying the suggested condition on the distance. If this condition is achieved, then remove the points of the constructed pair from V and put this pair in M (the output set of the solution). Then, choose a point and the nearest point of it from the remaining points in V to construct a pair and put it in M . Remove the two points of the constructed pair from V and repeat this process until V becomes an empty set. Obviously, this method is very simple. Furthermore, our algorithm can be applied without any modification on complete weighted graphs K_m and complete weighted bipartite graphs $K_{n,n}$, where $n, m \geq 1$ and m is an even.

Keywords- Perfect matching; approximation algorithm; Monte-Carlo technique; a randomized fully polynomial-time approximation scheme; and randomized algorithm.

I. INTRODUCTION

In this paper, the authors deal with Euclidean min-weighted perfect matching problem. This problem and its special cases are very important since they have several applications in many fields such as operations research, pattern recognition, shape matching, statistics, and VLSI, see [1], [2], and [3].

The previous studies treated the underlining problem in several versions, e.g. (un-)weighted general graphs, bipartite graphs, and a case of a set of points in Euclidean plane. For un-weighted bipartite graphs, Hopcroft and Karp showed that maximum-cardinality matchings can be computed in $O(m\sqrt{n})$

time [1]. In [4], Micali and Vazirani introduced an $O(m\sqrt{n})$ algorithm for computing maximum-cardinality matchings on un-weighted graphs. Goel et al. [5] presented an $O(n \log n)$ algorithm for computing the perfect matchings on un-weighted regular bipartite graphs.

In 1955, Kuhn used the Hungarian method for solving the assignment problem. He introduced the first polynomial time algorithm on weighted bipartite graphs having n vertices for computing min-weighted perfect matching in $O(n^3)$ time [6]. The matching problem on complete weighted graphs with $2n$ vertices was solved in $O(n^4)$ by Edmonds' algorithm [7]. In [8], Gabow improved Edmonds' algorithm to achieve $O(n(m + n \log n))$ time, where m is the number of edges in a graph. For the Euclidean versions of the matching problem, Vaidya [9] showed that geometry can be exploited to get algorithms running in $O(n^{5/2} \log^{O(1)} n)$ time for both the bipartite and nonbipartite versions. Agarwal et al. [10] improved this running time for the bipartite case to $O(n^{2+\delta})$, where $\delta > 0$ is an arbitrarily small constant. For nonbipartite case, the running time was improved to $O(n^{3/2} \text{poly} \log n)$, see details in [11].

Several researchers did more efforts to find good approximation algorithms for optimal matching which are faster and simpler than the algorithms obtained the optimal solutions (e.g. [12], [13]).

In [13], Mirjam and Roger gave an approximation algorithm for constructing a minimum-cost perfect matching on complete graphs whose cost functions satisfy the triangle inequality. The running time of that algorithm is $O(n^2 \log n)$ and its approximations ratio is $\log n$. For bipartite version on two disjoint n -point sets in the plane, Agarwal and Varadarajan [14] showed that an ε -approximate matching can be computed in $O((n/\varepsilon)^{3/2} \log^5 n)$ time. In [15], Agarwal and Varadarajan proposed a Monte Carlo algorithm for computing an $O(\log(1/\varepsilon))$ -approximate matching in $O(n^{1+\varepsilon})$ time. Based on Agarwal's ideas, Indyk [16] presented an $O(n \log^{O(1)} n)$ algorithm with probability at least $1/2$ and a cost at most $O(1)$ times the cost of the optimal matching. Sharathkumar and Agarwal [17] introduced a Monte-Carlo

algorithm that computes an ε -approximate matching in $O((n/\varepsilon^{O(d)}))poly \log n$ time with high probability.

In this paper, a randomized approximation algorithm for Euclidean min-weighted perfect matching problem is demonstrated. This algorithm follows Monte-Carlo technique. It computes an ε -approximate Euclidean matching of a set of $2n$ points with probability at least $1/2$. The probability is improved when running the algorithm more and more. We show that the algorithm is a *RFPTAS* (Randomized Fully Polynomial Time Approximation Scheme) for underline problem. We apply this algorithm for the Euclidean cases having $2n$ points and for general complete weighted graphs having $2n$ vertices.

The paper is organized as follows. In the next section, some basic definitions and concepts needed for the description of the topic are introduced. In section III, the description of our randomized algorithm for the min-weighted perfect matching problem in the plane is given. It is also proven that the given algorithm is a *RFPTAS*. Moreover, the same section contains some results of computational experiments on some classes of graphs. Finally, a conclusion of the paper is introduced in section IV.

II. BASIC DEFINITIONS AND CONCEPTS

Henceforth, let V be a set of $2n$ points in the plane. A matching of V is a collection M of pairs of V such that no point in V belongs to more than one pair in M [14]. A perfect matching of V is a matching of V in which every point in V belongs to exactly one pair of $z = d(u_1, u_7)$ and so, a perfect matching of V has n pairs. Let $d(u, v)$ be referred to the specified distance between u and v . The weight of a matching M is defined by summing the Euclidean distances between the paired points. For $\varepsilon > 0$, an ε -approximate perfect matching is a perfect matching M such that the weight of M is at most $(1 + \varepsilon)$ times the weight of a minimum weighted perfect matching [17]. In this paper, we consider the approximate Euclidean min-weighted perfect matching problem that finds an ε -approximate perfect matching of V . Let U be an optimization problem. An algorithm A is called a Randomized Fully Polynomial-Time Approximation Scheme (RFPTAS) [18] for U if there exists a function $p: N \times R^+ \rightarrow N$, such that for every input (x, ε) ,

$\Pr(A(x, \varepsilon) \in M(x)) = 1$ {for every random choice A computes a feasible solution of U },

$\Pr(R(x, \varepsilon) \leq 1 + \varepsilon) \geq 1/2$ {a feasible solution, whose approximation ratio is at most $(1 + \varepsilon)$, is produced with the probability at least $1/2$ }, and

$Time(x, \varepsilon^{-1}) \leq p(|x|, \varepsilon^{-1})$ and P is a polynomial in both its arguments $|x|$ and ε^{-1} .

In the approximate Euclidean bipartite min-weighted perfect matching problem, a real number $\varepsilon > 0$ and two disjoint sets V_1 and V_2 such that each of which has n points are given. In this version, if a pair (u, v) belongs to the matching, then $u \in V_1$ and $v \in V_2$. The solution of this problem is to find an ε -approximate perfect bipartite matching of $V_1 \cup V_2$.

In the following, the definition of an admissible pair is given. This definition is very important to design the main condition of our algorithm. We introduce first the meaning of a near point. Suppose V is a set of points. For any $v \in V$, $u \in V$ is called a near point of v if for all $u' \in V$; $d(v, u) \leq d(v, u')$.

Definition 2.1 Let V be a set of points. For any $u, v \in V$, a pair (u, v) is called an admissible pair if $d(u, v) \leq (1 + \varepsilon)d(u, u')$ or $d(u, v) \leq (1 + \varepsilon)d(v, v')$, otherwise it is called an inadmissible pair, where u', v' are the near points of u, v , respectively.

Let V be a set of points and $|V| = 2n$. For each pair $e = (u, v)$ and $u, v \in V$, we consider a random variable X_e defined by $X_e = 1$ if e is an inadmissible pair and $X_e = 0$ otherwise. Since each point $u \in V$ has at least a point $u' \in V$ such that the pair $e = (u, u')$ is an admissible pair (e.g. u' is the near point of u). Thus, $\Pr(X_e = 1) \leq (2n - 2)^2 / (2n)^2 = (n - 1)^2 / n^2$ and $\Pr(X_e = 0) \geq 1 - (n - 1)^2 / n^2$. Let X be a random variable defined by $X = \sum_{e \in M} X_e$, where M is any matching of V .

This means that X counts the number of inadmissible pairs in any matching M .

Let $v \in V$ and $V' \subset V$. Assume that v is a near point of every point in V' . The next lemma gives the upper bound of the cardinality of V' .

Lemma 2.1 Let V be a set of $2n$ points in the plane. If $v \in V$ and $V' \subset V$ satisfying v is a near point to every point in V' , then $|V'| \leq 6$.

Proof. Assume that $|V'| \geq 7$ and v is a near point to every point in V' . First, consider $|V'| = 7$, i.e., $V' = \{u_1, u_2, \dots, u_7\}$. In the plane, connect the point v to all points in V' as shown in Fig.1. Clearly from computational geometry that there exist at least one angle, α less than 60° . Consider w.l.o.g.

$\alpha = \angle u_1 v u_7$. To complete the proof, connect u_1 to u_7 . From Fig.1, $x = d(v, u_1)$, $y = d(v, u_7)$, and $z = d(u_1, u_7)$.

Note that $x \leq y$ So,

$$z^2 = (x \sin \alpha)^2 + (y - x \cos \alpha)^2 = x^2 + y^2 - 2xy \cos \alpha \quad \square$$

Since $\alpha < 60^\circ$ and $x \leq y$. So, $\cos \alpha > 1/2$ and we obtain $z < y$. This means that u_1 is a near one to u_7 rather than v and this is contradiction. Evidentially, when $|V'| > 7$, there exist more angles less than 60° and so we reach to the same contradiction. Therefore, cardinality of V' should be less than or equal 6. ■

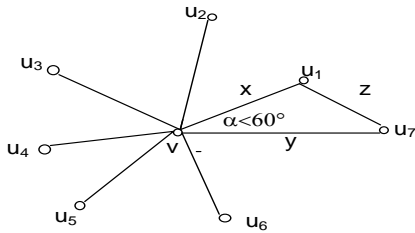


Figure 1. An illustrative example of the proof of Lemma 2.1.

III. A RANDOMIZED ALGORITHM FOR WEIGHTED PERFECT MATCHING

In this section, we describe a Randomized Perfect Matching, *RPM*, algorithm to construct an ϵ -approximate matching of a set V of $2n$ points in the plane. The analysis of running time and the derivation of an approximation ratio \square for this algorithm is introduced. We deduce that this algorithm is a Randomized Fully Polynomial Time Approximation Scheme.

This section is organized as follows. Section A is devoted to describe the *RPM*-algorithm. In addition, we derive that the *RPM*-algorithm is a RFPTAS for Euclidean min-weighted perfect matching problem in section B. Some results of computational experiments in cases of the set of $2n$ points in the plane, the Euclidean bipartite version, complete weighted graphs, and complete weighted bipartite graphs are contained in section C.

A. The description of *RPM*-algorithm

This part contains the details of the *RPM*-algorithm for Euclidean min-weighted perfect matching problem. The idea of the algorithm summarizes as follows. First, we choose uniformly at random a point v from V and get a near point to it from V , see u . Then, check the suggested condition of Definition 2.1 on $d(v, u)$: If the condition is satisfied, then the pair (v, u) is taking into the output solution M and eliminate v, u from V . Second, after repeating the process $\lceil 1/\epsilon \rceil$ times, choose uniformly at random a point u' , get a near point to it from V , see v' , and put a pair (u', v') in M .

Delete u', v' from V and repeat this work until V becomes empty. Note that deciding that a pair is admissible based on computing near points w.r.t. all $2n$ points in V ; i.e. before removing any point during the execution of the *RPM*-algorithm.

The steps of the *RPM*-algorithm are introduced in the following:

Algorithm *RPM*

Input: a set V of $2n$ points in the plane and a real number $\epsilon > 0$.

Output: M // a perfect matching of V .

N // the number of admissible pairs which is used insure that the probability of success $\geq 1/2$.

Begin

- 1: $M \leftarrow \varnothing$;
- 2: $k \leftarrow \lceil 1/\epsilon \rceil$;
- 3: $N \leftarrow 0$;
- 4: $W \leftarrow V$;
- 5: for $i = 1$ to k do
- 6: choose uniformly at random a point v from W ;
- 7: find u which is a near point to v in W ;
- 8: find v' which is a near point to v in V ;
- 9: find u' which is a near point to u in V ;
- 10: if a pair (v, u) is an admissible pair then
- 11: put a pair (v, u) in M ;
- 12: $W \leftarrow W - \{v, u\}$;
- 13: $N \leftarrow N + 1$;
- 14: while $W \neq \varnothing$ do
- 15: choose uniformly at random a point u from W ;
- 16: find v which is a near point to u in W ;
- 17: $W \leftarrow W - \{u, v\}$;
- 18: put a pair (u, v) in M ;
- 19: find v' which is a near point to v in V ;
- 20: find u' which is a near point to u in V ;

- 21: if a pair (u, v) is an admissible pair then
 22: $N \leftarrow N + 1$;
 23: return M and N ;
 End

As shown in the RPM-algorithm, the steps of this algorithm are simple and can be easily implemented. The main advantage of this method is that it can be applied on any weighted complete graph. This is because V can be considered as the set of vertices of a graph and the distance between any two vertices is measurable by the weight of an edge. Without any modification, this algorithm is also applied on any complete weighted bipartite graph. The results of the RPM-algorithm for solving the underlining problem in cases of the set of $2n$ points in the plane, the Euclidean bipartite version, complete weighted graphs, and complete weighted bipartite graphs are given in section C.

B. The analysis of RPM-algorithm

In this part, we prove that the RPM-algorithm is a RFPTAS for Euclidean min-weighted perfect matching problem. At first, we prove that the probability of success is greater than $1/2$, via using the following two lemmas.

Assume that M' is the set of inadmissible pairs obtained from the output matching of the RPM-algorithm. The following lemma gives an upper bound of the cardinality of M' .

Lemma 3.1 Let V be a set of $2n, n > 4$, points in the plane. If M' is the set of inadmissible pairs obtained from the produced matching M of the RPM-algorithm, then $|M'| \leq 5n/7$.

Proof. We partition the set V into disjoint $k, k > 1$, subsets such that each subset contains a point v and all points that v is a near point to each of them. From Lemma 2.1, we conclude that the number of subsets in this partition is at least $2n/7$. We can obtain at least an admissible pair in each subset in this partition. Thus, the number of admissible pairs in the matching M is at least $2n/7$. Since the cardinality of the matching M is n . Therefore, the number of inadmissible pairs in M is at most $5n/7$. ■

Lemma 3.2 Let V be a set of $2n$ points in the plane. Suppose M is a matching of V which is constructed by the RPM-algorithm and X is a random variable whose value is the number of inadmissible pairs in M . Then, $\Pr(X < n/2) > 1/2$.

Proof. Let M' be the set of inadmissible pairs in M . As mentioned in Lemma 3.1, $|M'| \leq 5n/7$. So, the number of points in inadmissible pairs is at most $10n/7$ and the probability of choosing one of them is at least $7/10n$. In the RPM-algorithm, the steps of for-loop repeat k times,

$k = \lceil 1/\varepsilon \rceil$, so the number of choosing points that do not give admissible pairs is at most $(7/10n)k$. Thus, the number of the admissible pairs which can be obtained in this loop is at least $((10n - 7)/10n)k$. Therefore, the number of the pairs that can be constructed in while-loop does not exceed $A = n - ((10n - 7)/10n)k$.

From linearity of expectation, we obtain

$$E(X) = E\left(\sum_{e \in M} X_e\right) = \sum_{e \in M} E(X_e) = \sum_{e \in M} \Pr(X_e = 1) \leq \frac{(n-1)^2}{n^2} A,$$

where X and X_e are given in section II.

Let $k > 30n^2 / (40n - 28)$, (as a result of executing of the RPM-algorithm more and more). From Morkov's inequality [19], $\Pr(X \geq n/2) \leq E(X)/(n/2) = (2(n-1)^2 A)/n^3$. By using $A = n - ((10n - 7)/10n)k$,

$$\Pr(X \geq n/2) \leq 2(n-1)^2 / n^3 (n - ((10n - 7)/10n)k).$$

Substituting $k = 30n^2 / (40n - 28)$, we get $\Pr(X \geq n/2) < (n-1)^2 / 2n^2 < 1/2$.

Therefore, $\Pr(X < n/2) = 1 - \Pr(X \geq n/2) > 1/2$. This completes the proof. ■

Now, we show that the RPM-algorithm is a RFPTAS.

Theorem 3.3 The RPM-algorithm is a RFPTAS for Euclidean min-weighted perfect matching problem.

Proof. First, we determine the running time of the RPM-algorithm. The time of steps 7- 9 is proportional to $O(n)$. So, the for-loop is executed in time $O(nk)$, where $k = \lceil 1/\varepsilon \rceil$.

Since the number of the pairs that can be constructed in while-loop does not exceed $A = n - ((10n - 7)/10n)k$ and the time of steps 16 and 19- 20 are proportional to $O(n)$, so the time of while-loop is $O(nA)$. Thus, the running time of the RPM-algorithm is $O(nk + nA)$.

Now, we show that the RPM-algorithm finds an ε -approximate perfect matching of V with probability at least $1/2$. It is clear that the output solution M of the RPM-algorithm is a perfect matching of V with cardinality n . From Lemma 3.2, we obtain that the number of inadmissible pairs in M is less than $n/2$ with probability at least $1/2$, hence the number of admissible pairs in M is greater than $n/2$ with the same probability. Let m be the number of admissible pairs in M . According to Definition 2.1, any admissible pair (u, v) satisfies $d(u, v) \leq (1 + \varepsilon)d(u, x)$ or $d(u, v) \leq (1 + \varepsilon)d(v, y)$, where x is a near point of u and y is a near point of v in V . Without loss of generality, we consider the pair (u, x) satisfies the above inequality. So,

$\sum_m d(u,v) \sum_m d(u,x)$. Since u is one coordinate of a pair in an optimal solution, OPT, see (u,z) and $m > n/2$ with probability at least $1/2$. Thus,

$$\sum_m d(u,v) \leq (1 + \varepsilon) \sum_m d(u,z) < (1 + \varepsilon) OPT$$

Obviously, one can deduce from the last equation that the approximation ratio of the RPM-algorithm is at most $(1 + \varepsilon)$ with probability at least $1/2$. This means that the RPM-algorithm is a RFPTAS. ■

C. The results of RPM-algorithm

In this section, we provide some results of the experimental evaluation of the proposed algorithm. We will show that the RPM-algorithm can be applied on four classes without any modification in its steps. These classes are:

- Class 1: Euclidean plane.
- Class 2: Euclidean bipartite plane.
- Class 3: complete weighted graphs.
- Class 4: complete weighted bipartite graphs.

In the first two classes, the inputs of the RPM-algorithm are a real number $\varepsilon > 0$ and a set V . V is the set of $2n$ points which are generated at random. The inputs of the RPM-algorithm in the last two classes are a real number $\varepsilon > 0$ and a complete weighted (bipartite) graph. The weights of edges are positive numbers which are generated at random (in Classes 3, 4) and are the Euclidean distances between any two points (in Classes 1, 2). In table 1, we introduce the numbers of admissible pairs which are produced from the RPM-algorithm for some examples of each class.

Table 1. The performance of the RPM-algorithm.

The values of n	Class 1	Class 2	Class 3	Class 4
20	13	18	16	19
50	33	48	36	49
100	72	90	66	94
300	215	273	190	283
500	364	450	295	454
600	436	538	364	544
1000	714	910	577	912
1200	848	1068	865	1088
1500	1044	1351	873	1379
2000	1376	1776	1122	1819
2030	1395	1807	1137	1850

Since the number of admissible pairs is greater than $n/2$ for each example, so the probability of success of the RPM-algorithm is greater than $1/2$. From these results, the algorithm is suitable for execution on all examples and for any value of n .

Fig. 2 shows the relation between the values of n and the number of admissible pairs. The algorithm is implemented by the Java language. The program works on a 3.00GHz Pentium IV personal computer.

IV. CONCLUSION

In this paper, we introduce a randomized approximation algorithm which follows a Monte-Carlo method for treating the weighted perfect matching problem in the plane. The idea of this algorithm is very simple. Since this algorithm succeeds in computing an ε -approximation matching with probability at least $1/2$ in $O(nk + nA)$ time, where $k = \lceil 1/\varepsilon \rceil$ and $A = n - ((10n - 7)/10n)k$, so this algorithm is a RFPTAS. Our algorithm can be applied without any modification on complete weighted graphs K_m and complete weighted bipartite graphs $K_{n,n}$, where $n, m \geq 1$ and $m = 2n$. The inputs of our algorithm are a set V of $2n$ points in the plane or a complete weighted graph and a real number $\varepsilon > 0$ and the outputs are a perfect matching of an input instance and the number of admissible pairs in this matching. The numbers of admissible pairs which are produced from the suggested algorithm for some examples of each state are summarized in Table 1.

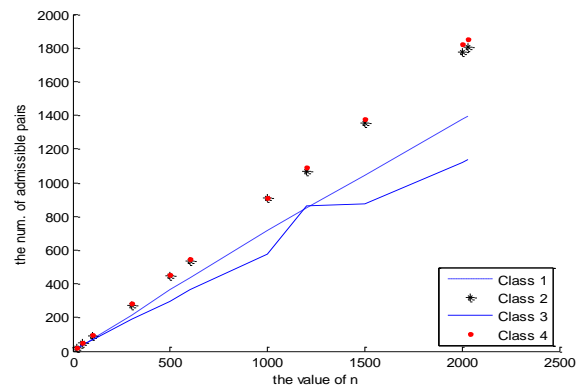


Figure 2. The relation between the values of n and the number of admissible pairs.

REFERENCES

- [1] J. E. Hopcroft and R. M. Karp, An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs, SIAM J. Comput., (1973) 2(4):225–231.
- [2] L. Lovász, On determinants, matchings and random algorithms, In L. Budach, editor, Fundamentals of Computation Theory, FCT '79, Akademie-Verlag Berlin (1979), pages 565–574.
- [3] O. Marcotte and S. Suri, Fast matching algorithms for points on a polygon, In Proc. 30th Annu. IEEE Sympos. Found. Comput. Sci., 1989, pages 60–65.
- [4] S. Micali and V. V. Vazirani, An $o(\sqrt{v})$ algorithm for finding maximum matching in general graphs, In FOCS, 1980, pages 17–27.
- [5] A. Goel, M. Kapralov, and S. Khanna, Perfect matchings in $O(n \log n)$ time in regular bipartite graphs, In STOC, 2010, pages 39–46.
- [6] H. W. Kuhn, The Hungarian method for the assignment problem, Naval Research Logistics Quarterly, 1955, 2:83–97.
- [7] J. Edmonds, Paths, trees and flowers, Canadian Journal of Mathematics, 1965, 17:449–467.

- [8] H.N. Gabow, Data structures for weighted matching and nearest common ancestors with linking, In Proceedings of the 1st Annual ACM-SIAM Symposium on Discrete Algorithms, 1990, pages 434–443.
- [9] P. M. Vaidya, Geometry helps in matching. SIAM J. Comput., 1989, 18:1201–1225, December.
- [10] P. K. Agarwal, A. Efrat, and M. Sharir, Vertical decomposition of shallow levels in 3-dimensional arrangements and its applications, SIAM J. Comput., 1999, 29:39–50.
- [11] K. R. Varadarajan, A divide-and-conquer algorithm for min-cost perfect matching in the plane, In FOCS, 1998, pages 320–331.
- [12] M.X. Goemans and D.P. Williamson, A general approximation technique for constrained forest problems, In SODA: ACM-SIAM Symposium on Discrete Algorithms (A Conference on Theoretical and Experimental Analysis of Discrete Algorithms), 1992.
- [13] M. Wattenhofer and R. Wattenhofer, Fast and Simple Algorithms for Weighted Perfect Matching, Electronic Notes in Discrete Mathematics 17 (2004) 285–291.
- [14] K.R. Varadarajan and P.K. Agarwal, Approximation algorithms for bipartite and non-bipartite matching in the plane, In SODA: ACM-SIAM Symposium on Discrete Algorithms (A Conference on Theoretical and Experimental Analysis of Discrete Algorithms), 1999.
- [15] P. K. Agarwal and K. R. Varadarajan, A near-linear constant-factor approximation for Euclidean bipartite matching, In Proc. of 12th Annual Sympos. on Comput. Geom., 2004, pages 247–252.
- [16] P. Indyk, A near linear time constant factor approximation for Euclidean bichromatic matching (cost), In Proc. Annual Sympos. on Discrete Algorithms, 2007, pages 39–42.
- [17] R. Sharathkumar and P. K. Agarwal, A Near-Linear Time ϵ -Approximation Algorithm for Bipartite Geometric Matching, Accepted, to appear in STOC 2012.
- [18] J. Hromkovi, Algorithmics for hard problems, Springer, 2001.
- [19] S. M. Ross, Introduction to Probability Models, Academic press an imprint of Elsevier, Vol. 9, 2007, pp. 77-78.
- [20] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (*references*)
- [21] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [22] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
- [23] K. Elissa, “Title of paper if known,” unpublished.
- [24] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
- [25] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetism Japan, p. 301, 1982].
- [26] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

Cost Analysis of Algorithm Based Billboard Manger Based Handover Method in LEO satellite Networks

Suman Kumar Sikdar

Department of Computer science & Engineering,
University of Kalyani
Kalyani-741235, India

Soumaya Das

Department of E.T.C.E Bengal Institute of technology,
West Bengal University of technology
Kolkata, India

DebabrataSarddar

(Asth Professor) Department of Computer science
& Engineering, University of Kalyani,
Kalyani-741235, India

Abstract—Now-a-days LEO satellites have an important role in global communication system. They have some advantages like low power requirement and low end-to-end delay, more efficient frequency spectrum utilization between satellites and spot beams over GEO and MEO. So in future they can be used as a replacement of modern terrestrial wireless networks. But the handover occurrence is more due to the speed of the LEOs. Different protocol has been proposed for a successful handover among which BMBHO is more efficient. But it had a problem during the selection of the mobile node during handover. In our previous work we have proposed an algorithm so that the connection can be established easily with the appropriate satellite. In this paper we will evaluate the mobility management cost of Algorithm based Billboard Manager Based Handover method (BMBHO). A simulation result shows that the cost is lower than the cost of Mobile IP of SeaHO-LEO and PatHO-LEO.

Keywords-Component; Handover latency; LEO satellite; Mobile Node (MN); Billboard Manager (BM).

I. INTRODUCTION

Modern terrestrial networks are designed as per to give the low cost and best quality service. Mobile networks provide communication to a limited geographical area. The applications of satellite network increases in order to provide the global coverage in a large area. There are different satellites for this communication [1] [2].

1. Geostationary satellite
2. Medium Earth Orbit satellite
3. Low Earth Orbit satellite.

Among which Low Earth Orbit satellite is the best for communication and as the replacement of future terrestrial wireless networks as it has some advantages like

- a. Low propagation delay
- b. Low end to end delay
- c. Low power requirement

d. More efficient frequency spectrum utilization between satellites and spot-beams.

But the Leo satellite has also some problems. The main problem is that the low earth orbit satellites have a large relative speed than the speed of mobile nodes (MN) & earth. That's why the handover occurrence is more. So the call blocking probability (P_b) and force call termination probability (P_f) is also higher. To solve this problem different handover techniques have been proposed.

What is handover?

Handover is the process of transferring satellite control responsibility from one earth station to another earth station without any loss or interruption of the service[3][4].

An unsuccessful handover can degrade the system performance like call quality as well as it call because the forced call termination. To solve these problems different handover technique has been proposed [5] [6]. A handover is done in the following three steps:

- i. Scanning
- ii. Authentication
- iii. Re-association

Scanning: When a mobile station is moving away from its current satellite, it initiates the handoff process when the received signal strength and signal-to-noise-ratio have decreased below the threshold level. The MN now begins the scanning to find new satellite. It can either go for a passive scan (where it listens for beacon frames periodically sent out by satellites) or choose a faster active scanning mechanism wherein it regularly sends out probe request frames and waits for responses for T_{MIN} (min Channel Time) and continues scanning until T_{MAX} (max Channel Time) if at least one response has been heard within T_{MIN} . Thus, $n * T_{MIN} \leq$ time to scan n channels $\leq n * T_{MAX}$. The information gathered is then processed so that the MN can decide which Satellite to join next. The total time required until this point constitutes 90% of the handoff delay.

Authentication: To associate the link with the new satellite Authentication is necessary. Authentication must either immediately proceed to association or must immediately follow a channel scan cycle. In pre-authentication schemes, the MN authenticates with the new satellite immediately after the scan cycle finishes.

Re-Association: Re-association is a process for transferring associations from old satellite to new one. Once the MN has been authenticated with the new satellite, re-association can be started. Previous works has shown re-association delay to be around 1-2 ms. The range of scanning delay is given by:-

$$N \times T_{min} + T_{scan} + N \times T_{max}$$

Where N is the total number of channels according to the spectrum released by a country, T_{min} is Min Channel Time, T_{scan} is the total measured scanning delay, and T_{max} is Max Channel Time. Here we focus on reducing the scanning delay by minimizing the total number of scans performed.

The paper is organized as follows: In the second section we have described the related works on handover management including Mobile IP, SeaHO-LEO, PatHO-LEO and our previously proposed Algorithm Based BMBHO. In the third section we have described the proposed work which includes the cost analysis of Algorithm Based BMBHO. In the fourth section the simulation results of both our method and standard methods. In the section 5 we conclude the whole paper and finally a future work is mention regarding this paper in section six.

II. RELATED WORK:

To solve the problems of unsuccessful handovers different handover management protocols have been proposed [7]. The most widely used one is MIP i.e. Mobile IP Network [8]. It is proposed by The Internet engineering task force (IETF) to handle mobility of internet hosts for mobile data communications. MIP is based over the concept of Home Agent (HA) and Foreign Agent (FA) for delivering of packets from one MN to CN.

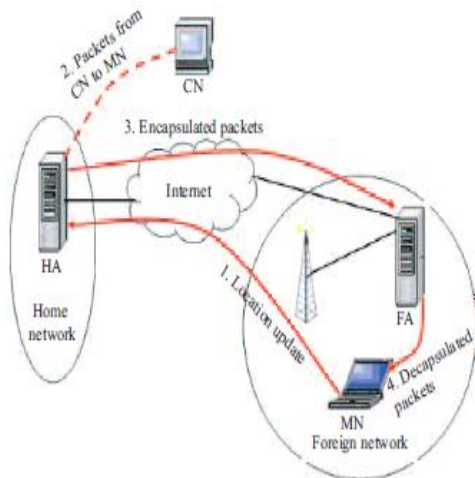


Figure 1: Handover scenario in MIP It is done in four steps.

- At the beginning of the handover MN registers itself in FA and waits for the channel allocation in FA and also updates its location in HA directory.
- Then the packets are sent to HA and HA encapsulate it.
- After that encapsulated packets are sent to The FA.
- Lastly FA decapsulate those packets and sent it to MN.

But this protocol also has some drawbacks. The main drawbacks of this protocol are:

- It has High handover latency
- packet lost rate is also very high
- It has insufficient routing path
- Conflicts with network security solution

So to overcome this drawbacks another protocol have been proposed i.e. Seamless handover management scheme (SEAHO-LEO) [9] [10] [11].

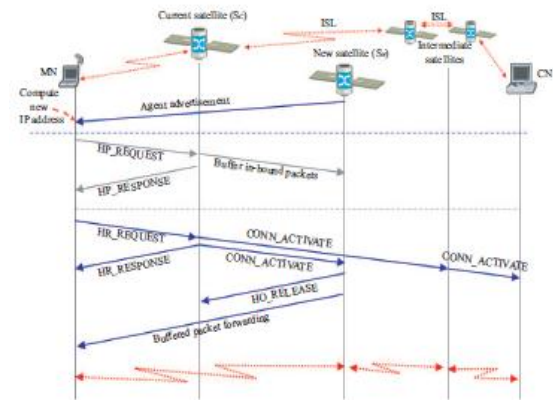


Figure 2: Handover scenario in SeaHO-LEO

This can be done in the following steps:

- Calculate a new IP
- Send handover preparation request to current satellite
- Start to use new IP to send data packets
- CN starts to use new satellite

The main disadvantage of this process is

- packet loss is less
- It has lower handover latency.

The main disadvantage of this process is

- High messaging traffic.

To get over these drawbacks another method to remove high messaging traffic is Pattern based handover management (PatHO-LEO) [8], [9].

It describes as follows

- Satellite register to BM.
- MN registers to BM.
- BM establishes the satellite and user mobility pattern (SMUP) table.
- CN and BM establish connection.
- CN sends data packets to MN.

But the main drawback of PatHO-LEO is that

Every user should have a specific mobility pattern in a specific period of time. A user can have more than one mobility pattern. But when it violates its mobility pattern the handover process will be either in SeaHO-LEO or MIP.

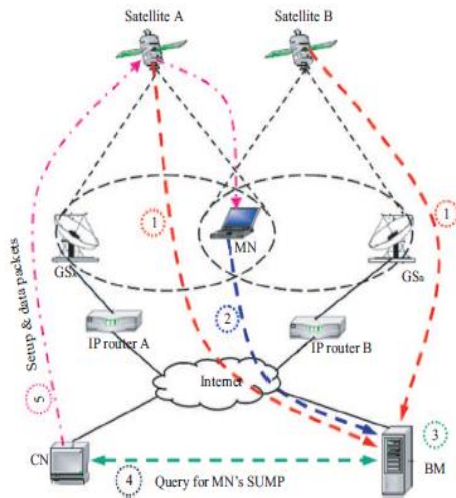


Figure 3: Handover scenario in PatHO-LEO

The no of user who do not have a specific mobility pattern in a week is increasing day by day like salesman, LIC worker who have to go different place at different time in a week.

Now to overcome these existing problems the handover management protocol proposed is named as BMBHO i.e. Billboard Manager Based Handover Technique where BM only stores satellite location signal strength based on QoS parameters[12][13].

❖ In BMBHO we assume that the direction of the signal flow is in one side or both side i.e. from CN to MN (where CN is fixed & MN is movable) or from MN1 to MN2 and vice versa (where both are movable).

❖ If the CN/MN2 is under the footprint of same satellite then the communication will be via one satellite otherwise via different satellite by ISL. So it is not important to know that communication is through CN/MN2, as the method is same for all.

This handover process can be done in the following ways

1) **BM stores all info about satellites:** the entire satellites register to BM including their IP address. This information not subjected to change and permanently stored in the BM database.

2) **All satellite sends periodic info:** All the satellites will send the following info periodically to the BM.

i) **Channel capacity:** -- How many channels are available in the satellite.

ii) **Signal strength:** --What is the strength of the signal at that time because from time to time and area to area due to the different weather condition.

This information is not constant & it updates itself every time it gets a new info. The time period of this update will be set as small as possible because a huge no of MN lies under

the footprint of a satellite. So the channel capacity changes very frequent. This time period is inversely proportional to the success of handover.

3) **Handover request is send to BM:** If a new MN wants to handover i.e. its signal strength decreases under a certain level called threshold level, it sends a HANDOVER_REQUEST (HO_REQ) to BM via its current satellite which contains the following

i) IP addresses of the current satellite (CS),

ii) IP address of adjacent satellite (AS) If MN/MN1 is connected to CN/MN2 through more than one satellite by ISLs.

iii) IP address of MN itself.

ii) Position of MN.

iii) The direction of the MN i.e. in which direction it wants to go.

4) **BM selects the new satellite:** Now BM first makes a list of available satellites in that direction at that time with the help of its stored data & the updates of satellites. Then BM selects best satellites for that MN according to the QoS parameters. A specific algorithm has to be developed for selecting the correct satellite.

5) **MN starts to use new satellite:** Once the satellite is selected BM sends the IP address of the new satellite to the MN & CN/MN2/AS. Now CN/MN2/AS makes a connection set up for the new satellite and it communicates to MN via the new satellite.

ALGORITHM:

The algorithm of BMBHO [14] is as follows

1) BM stores all information about satellite like IP addresses of the satellite.

2) All satellite sends periodic information to Billboard Manager.

a) Channel capacity

b) Signal strength

Both of the information varies time to time and also area to area.

3) Now for $t=0$, compare channel capacity if the channel capacity >0

Continue;

Else stop

4) Compare channel capacity, choose the maximum one.

5) If the channel capacity of the two satellite to handover is same,

6) Compare the signal strength. Choose the lowest signal strength of same channel capacity.

Else go back to 4

7) Repeat 4-6 every time while choosing a new satellite to handover.

8) Make a list of the available satellites and store it to BM

9) Now, If a new mobile node wants to handover, signal strength decreases under a certain level i.e. threshold level, it

sends a Handover Request to BM via its current satellite containing

- a) IP address of the current satellite.
- b) IP address of the adjacent satellite, If MN/MN1 is connected to CN/MN2 through more than one satellite by ISLs.
- c) IP address of MN.
- d) Position of MN
- e) The direction of the MN

10) Now BM again makes a list of available MNs.

11) Now comparing the first list and second list it chooses the best satellite to handover.

12) Once the satellite is selected, BM sends MN the IP address of the new satellite.

13) Now the connection is established.

FLOW CHART:

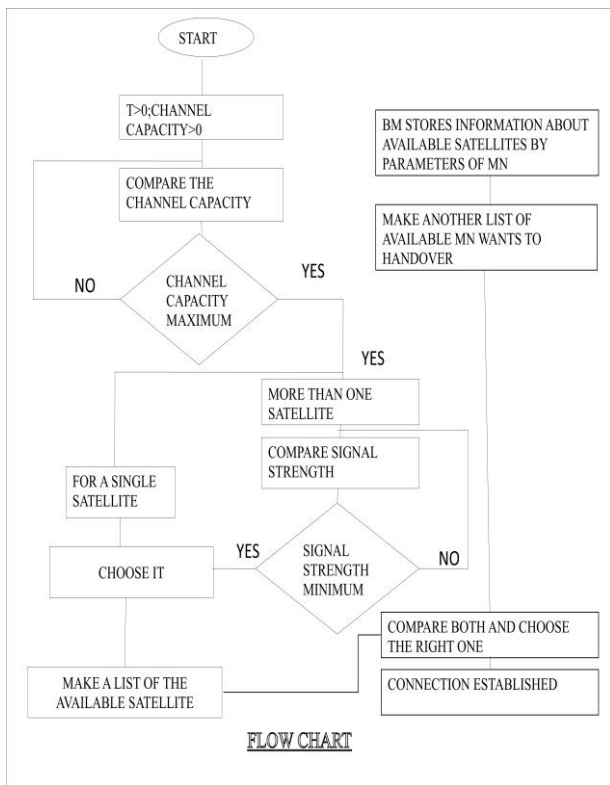


Figure 4: Algorithm for establishing connection

III. PROPOSED WORK:

In our previous work we have discussed the Algorithm based BMBHO method and shown how it can reduce handover latency and call blocking probability.

Here we will analysis the cost of this Algorithm based BMBHO method and compare it with other standard methods like MIP.

Mobility Management Cost Definition

In [12] the mobility management cost is evaluated as the product of generated control message size, M and the number of hopes, H, required to deliver the message. If we apply such definition into the paging cost, it will be proportional with the

number of receivers. Taking into account the broadcasting capabilities of satellites, however, the cost is also simply a product of the message size and the number of travelled hops.

$$\text{Cost} = M \cdot H \quad (1)$$

Costs of different Mobility management events:

The following defines the cost required for each mobility management event; binding update, local forwarding and paging

For each case, the Control messages generated are assumed to be equally sized (M) in all the four events.

The number of control messages that are generated upon a handover occurrence between mobile nodes and the corresponding ARs, is assumed to be same for MIP and our proposed method. Thus we can neglect the number of control message in the cost evaluation.

1. **Binding Update Cost:** Let $H_{MN,LD}$ denote the number of hops between a mobile node and the Location Directory. The cost for binding update procedure can be expressed as:

$$M \cdot H_{MN,LD}$$

2. **Local Forwarding Cost:** Denoting the number of hops between two adjacent satellites as $H_{AR,AR}$ the local forwarding cost is shown as follows:

$$M \cdot H_{AR,AR}$$

Management Cost of MIP and our proposed method

The costs of Mobile IP and our proposed method are as follows

A. **Mobile IP:** The cost of MIP is the product of binding update cost and rate of handover occurrence. The local forwarding, paging and GPS are not used here. So the MIP management cost, $C_{MIP}(t)$ can be expressed as

$$C_{MIP}(t) = M \cdot H_{MN,LD} \cdot R_{HO}(t) \quad (2)$$

Where the rate of handover occurrence, $R_{HO}(t)$, is:

$$R_{HO}(t) = V_{sat} \cdot L_{sat} \int_{V_{sat}(t-t\Delta)}^{V_{sat}t} DL(V_{sat}.t) dt \quad (3)$$

Where, V_{sat} and L_{sat} denote the ground speed of satellite and the coverage boundary length, respectively. $DL(V_{sat}.t)$ denote the linear density of nodes on the coverage boundary at time t .

B. **PatHO-LEO:** In the PatHO-LEO model, the local forwarding and paging scheme create some additional cost. The total cost of PatHO-LEO model $C_{PatHO-LEO}(t)$ is

$$C_{PatHO-LEO}(t) = M \cdot H_{MN,LD} + M \cdot H_{AR,AR} R_{HO}(t) \cdot \alpha + \{M \cdot H_{AR,AR} (S-1) + M \cdot S\} \cdot n(t) \cdot (1-\alpha) \cdot \lambda \quad (4)$$

Where, $H_{AR,AR}$ and S denote the number of hops between two adjacent satellites and the number of single-beam satellites that cover a single paging area, respectively. $n(t)$ and α denote the total number of MNs per a coverage area at time t and the ratio of active MNs to the total number of MNs, respectively. The rate of new connections to a MN is denoted as λ .

C. **Proposed Work:** In our proposed method we have introduced the billboard manager (BM) so that we have

successfully removed the scanning cost but it will introduce the messaging cost which we will evaluate now

1. Messaging Cost in a day: As all channel sends periodically the messages to BM which contains two information channel capacity and signal strength so if N be the total number of satellite and t (in sec) will be the time interval between sending two successive messages so the total number of messaging in a day will be

$$C_{MSG, DAY} = 24 * \{(M.H_{Sat, BM} * 3600) / t\} * n \quad (5)$$

2. Messaging Cost per handover: Now we will evaluate the messaging cost per handover. When the signal strength and signal to noise ratio decreases below the threshold level then every MN sends the handover request HR_REQ and after performing the BMBHO algorithm BM sends IP of the current satellite and adjacent satellite to MN so the required messaging cost between MN and BM $C_{MN, BM}$ is

$$C_{MN, BM} = 2 * M * H_{MN, BM} \quad (6)$$

Now the message transfer between satellites to BM is also two which contains the IP address of the MN; one to the current satellite and another to the adjacent satellite. So the total messaging cost between BM and satellites $C_{Sat, BM}$ is

$$C_{Sat, BM} = M * (H_{CSat, BM} + H_{AdSat, BM}) \quad (7)$$

Where $H_{CSat, BM}$ is the message transfer between the current satellite and BM and $H_{AdSat, BM}$ is the message transfer between adjacent satellite and BM.

Now if the handover involves only one satellite then

$$H_{AdSat, BM} = 0$$

Let K be the total number of handover in a day and L be the total number of handover which involves only one satellite then equation 7 reduces to

$$C_{Sat, BM} = M * \{H_{CSat, BM} * (K-L) + H_{AdSat, BM} * L\} \quad (8)$$

So now the total messaging cost for handover $C_{MSG, HO}$ is

$$\begin{aligned} C_{MSG, HO} &= C_{MN, BM} + C_{Sat, BM} \\ &= 2 * M * H_{MN, BM} + M * \{H_{CSat, BM} * (K-L) + H_{AdSat, BM} * L\} \end{aligned} \quad (9)$$

So the total messaging cost C_{MSG} is

$$\begin{aligned} C_{MSG} &= C_{MSG, DAY} + C_{MSG, HO} \\ &= 24 * \{(M.H_{Sat, BM} * 3600) / t\} * n + 2 * M * H_{MN, BM} \\ &\quad + M * \{H_{CSat, BM} * (K-L) + H_{AdSat, BM} * L\} \end{aligned} \quad (10)$$

So the total cost of handover C_{Tot} is

$$\begin{aligned} C_{Tot} &= (C_{MSG} + M.H_{AR, AR}) * R_{HO}(t) \\ &= \{(24 * \{(M.H_{Sat, BM} * 3600) / t\} * n + 2 * M * H_{MN, BM} \\ &\quad + M * \{H_{CSat, BM} * (K-L) + H_{AdSat, BM} * L\}) + M.H_{AR, AR}\} * R_{HO}(t) \end{aligned} \quad (11)$$

Equation 11 represents the total cost of algorithm based BMBHO.

IV. SIMULATION RESULT:

In order to evaluate the performance of the new algorithm based BMBHO, we compared it to MIP & SeaHO-LEO and the previous BMBHO. Each algorithm is evaluated by analyzing the Handoff delay, Forced call termination probability & MN's throughput and efficiency.

The simulation results were run on MATLAB 7.8 in a designed virtual environment.

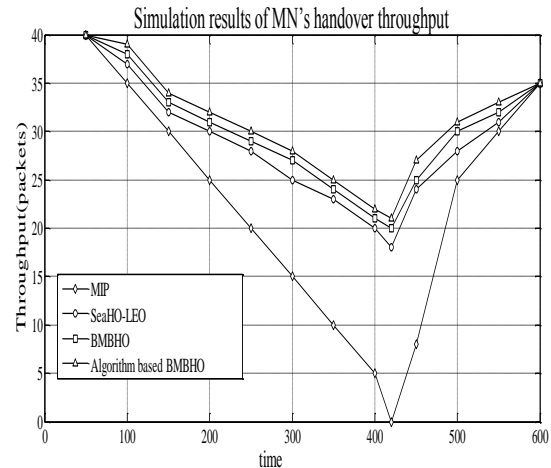


Fig 5: Simulation results of MN's handover throughput

In figure 5 we compare the Handover throughput for MIP, SeaHO-LEO & BMBHO and algorithm based BMBHO during a handover. Due to the tunneling between HA and FA in Mobile IP network, throughput of the channel between MN1/CN and MN2/MN converges to zero during handover and the handover model is completed, the throughput reaches a reasonable value. SeaHO-LEO throughput is better than MIP during handover as it does not reach to zero.

In BMBHO the throughput is higher than SeaHO-LEO because the handover takes very less time and the packets during handover is sent by the old link. And here we can see that algorithm based BMBHO is far better as it has a specific algorithm to choose the best satellite for establishing connection.

In figure 6 we have compared the handoff latency between the MIP, SeaHO-LEO, BMBHO and the Algorithm based BMBHO. Comparing all the results we can conclude that due to omitting the scanning process handoff delay is very less in BMBHO than the other two.

Now in the Algorithm Based BMBHO, we have taken the BMBHO with just an algorithm to select the satellite. So for this also handoff latency will be lesser than the MIP and SeaHO-LEO and as this finds the easiest way and lesser time to establish the connection following the specific algorithm its handoff latency is lesser than the normal BMBHO also.

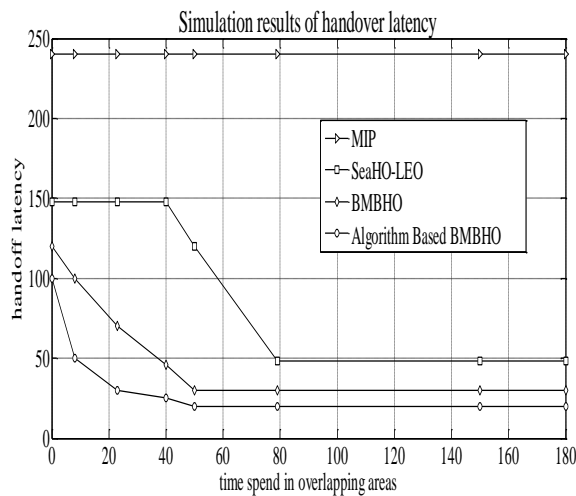


Fig 6: Simulation results of handover latency

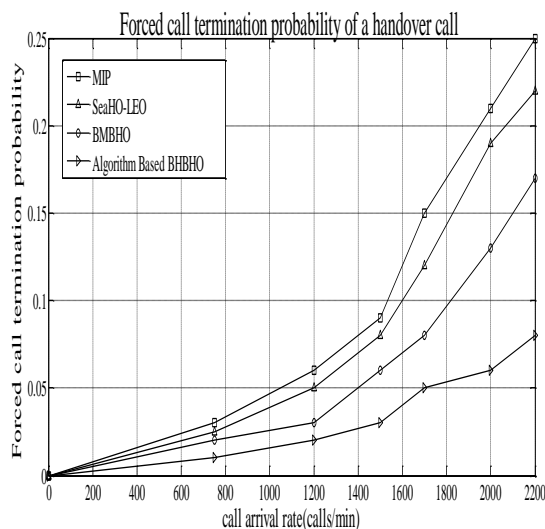


Figure7: Forced call termination probability of a handover call

In figure 7 we compare the Forced call termination probability of MIP & SeaHO-LEO, BMBHO with the Algorithm Based BMBHO. Among these management models, BMBHO has the lowest Forced call termination probability. In MIP there is a channel allocation so MN has to wait and if there is no free channel within the handoff time the call will be terminated. In SeaHO-LEO the MN has to wait for the agent advertisement from a new satellite. If it did not received within handoff time the call is being terminated. But in BMBHO the no of channel available in the satellites seen by the MN at the time of handoff is already known to BM so BM selects the new satellite for MN which has a free channel. So the force call termination probability is reduced. Now in our approach since there is a specific algorithm for choosing the satellite. It will choose the best satellite with maximum channel capacity and for that the connection will be efficient and so is the communication and data transfer. That is why there the call termination probability is almost equal to zero. By the above result we can show that for even 2200 calls per minute the forced call termination probability is almost equal to 0.07 whereas for MIP it is 0.25, for SeaHO-LEO it is 0.21 and for BMBHO it is 0.17.

V. ACKNOWLEDGMENT

In this paper we have evaluated the total cost analysis of algorithm based BMBHO management where we have shown that the specific algorithm can reduce handover latency, data loss, scanning time, cost and forced call termination probability as well as can increase the MN's throughput and the efficiency.

We first described the handover is and handover process. Then we described the standard handover mechanism MIP and also SeaHO-LEO and PathO-LEO and our BMBHO and their drawbacks. Then we have shown the specific algorithm to reduce the drawback of BMBHO. Then we have evaluated the total cost of Algorithm Based BMBHO. Relying on the simulation results we showed that our proposed mechanism reduced handoff latency and data transfer. This algorithm can help BM to choose the best satellite for handover so that the call quality increases as well as the call dropping probability reduces to zero. Our method is more efficient than the standard one to establish the best connection.

FUTURE WORK

In future we will be focused on how to reduce the cost of Algorithm Based BMBHO in LEO satellite Networks.

REFERENCES

- [1] S. L. Kota, P. A. Leppanen, and K. Pahlavan, *Broadband Satellite Communications For Internet Access*, Kluwer Academic Publishers, 2004.
- [2] A. Jamalipour, "Satellites in IP networks," in *Wiley Encyclopedia of Telecommunications*, vol. 4, Wiley, 2002, pp. 2111–2122.
- [3] Satellite Mobility Pattern Scheme for Central and Seamless Handover Management in LEO Satellite Networks Ays, eg'ulT'uys'uz and Fatih Alag'oz
- [4] H. Uzunalioglu, I. F. Akyildiz, Y. Yesha, and W. Yen, "Footprint handover
- [5] H. Uzunalioglu, I. F. Akyildiz, Y. Yesha, and W. Yen, "Footprint handover
Rerouting protocol for low earth orbit satellite networks," *Wireless Networks*, vol. 5, no. 5, pp. 327–337, 1999
- [6] Systems By Joydeep Banerjee D Sarddar, S.K. Saha, M.K. Naskar, T.Jana, U. Biswas
- [7] H. N. Nguyen, S. Lepaja, J. Schuringa, and H. R. Van As, "Handover management in low earth orbit satellite IP networks," *IEEE Global Telecommunications Conference*, San Antonio, TX, USA, pp. 2730–2734, 25–29 November 2001
- [8] J. T. Malinen and C. Williams, "Micromobility taxonomy," *Internet Draft*, IETF, Nov. 2001
- [9] T'uys'uz and F. Alag'oz, "Satellite mobility pattern based handover management algorithm in LEO satellites," in *Proc. IEEE ICC 2006*, Istanbul, Turkey, June 2006..
- [10] Ays, eg'ulT'uys'uz and Fatih Alag'oz, "Satellite Mobility Pattern Scheme for central and Seamless Handover Management in LEO Satellite Networks", *JOURNAL OF COMMUNICATIONS AND NETWORKS*, VOL. 8, NO. 4, DECEMBER 2006.
- [11] M. Atiquzzaman, S. Fu, and W. Ivancic, "TraSH-SN: A transport layer seamless handoff scheme for space networks," in *Proc. ESTC 2004*, Palo Alto, CA, June 2004.
- [12] Debabrata Sarddar, Soumya Das, Dipsikha Ganguly, Sougata Chakraborty, M.k.Naskar, A New Method for Fast and Low Cost Handover in Leo Satellites(*International Journal of Computer Applications (0975 – 8887) Volume 37– No.7, January 2012*) <http://www.ijcaonline.org/archives/volume37/number7/4622-6631>

- [13] Debabrata Sarddar, Soumya Das, Dipsikha Ganguly, Sougata Chakraborty, M.k.Naskar, A New Method for Controlling Mobility Management Cost of PatHO- LEO Satellite and Mobile IP Network(*International Journal of Computer Applications* (0975 – 8887)Volume37–No.7,January2012
<http://www.ijcaonline.org/archives/volume37/number7/4621-6630>).
- [14] Dipsikha Ganguly, Debabrata Sarddar, Soumya Das, **Suman Kumar Sikdar**, Sougata Chakraborty and Kunal Hui. Article: Algorithm Based Approach for the Connection Establishment in the Fast Handover in Leo Satellites in BMBHO. *International Journal of Computer Applications* 44(12):36-42, April 2012. Published by Foundation of Computer Science, New York, USA
- [15] Debabrata Sarddar, Dipsikha Ganguly, Soumya Das, **Suman Kumar Sikdar**, Sougata Chakraborty, Kunal Hui, Shabnam Bandyopadhyay, Kalyan Kumar Das and Sujoy Palit. Article: Cost Analysis of Location Manager based Handover Method for LEO Satellite Networks. *International Journal of Computer Applications* 45(19):1-6, May 2012. Published by Foundation of Computer Science, New York, USA
- [16] P. Bhagwat, C. Perkins, and S. Tripathi, “Network layer mobility: An architecture and survey,” *IEEE Pers. Commun.*, vol. 3, no. 3, pp. 54–64, June1996.
- [17] A. T. Campbell, J. Gomez, S. Kim, Z. Turanyi, C.-Y. Wan, and A. Valko, Comparison of IP micro-mobility protocols,” *IEEE Wireless Commun. Mag.*, vol. 9, no. 1, Feb. 2002.
- [18] H. Tsunoda, K. Ohta, N. Kato, and Y. Nemoto, “Supporting IP/LEO satellite networks by handover-independent IP mobility management,” *IEEE J.Select. Areas Commun.*, vol. 22, no. 2, pp. 300–307, 2004.
- [19] Debabrata Sarddar, Dipsikha Ganguly, Soumya Das, **Suman Kumar Sikdar**, Sougata Chakraborty, Kunal Hui, Shabnam Bandyopadhyay, Kalyan Kumar Das and Sujoy Palit. Article: Cost Analysis of Mobile IP for the Repetitive IP Stations during a Short Period of Time. *International Journal of Computer Applications* 45(19):7-12, May 2012. Published by Foundation of Computer Science, New York, USA.
- [20] S. Kalyanasundaram, E.K.P Chong, and N.B. Shroff, “An efficient scheme to reduce handoff dropping in LEO satellite systems,” *Wireless Networks*, vol. 7, no. 1, pp. 75–85, January 2001.
- [21] H. N. Nguyen, S. Lepaja, J. Schuringa, and H. R. Van As, “Handover management in low earth orbit satellite IP networks,” *IEEE Global Telecommunications Conference*, San Antonio, TX, USA, pp. 2730–2734, 25-29 November 2001.

AUTHORS PROFILE

Suman Kumar Sikdar is currently pursuing his PhD at Kalyani University . He completed his M.Tech in CSE from jadavpur University in 2011 and B-Tech in Computer Science & Engineering from Mursidabad College of engineering and technology under West Bengal University of Technology in 2007. His research interest includes wireless communication and satellite communication. Email:sikdersuman@gmail.com

Soumya Das, son of Mr. Subrata Das and Mrs. Swapna Das, currently pursuing his B.Tech in Electronics & Communication Engg. at Bengal Institute of Technology under West Bengal University of Technology. His research interest includes mobile communication & satellite communication.

Debabrata Sarddar (Asst. Professor in Kalyani University) is currently pursuing his PhD at Jadavpur University. He completed his M.Tech in Computer Science & Engineering from DAVV, Indore in 2006, and his B.Tech in Computer Science & Engineering from Regional Engineering College(NIT), Durgapur in 2001. His research interest includes wireless and mobile communication Email:dsarddar@rediffmail.com

An agent based approach for simulating complex systems with spatial dynamics application in the land use planning

Fatimazahra BARRAMOU

Architecture System Team – LISER
Hassan II University- AinChock ENSEM Casablanca,
Morocco

Malika ADDOU

Architecture System Team – LISER
Hassania School of Public Works EHTP
Casablanca, Morocco

Abstract—In this research a new agent based approach for simulating complex systems with spatial dynamics is presented. We propose architecture based on coupling between two systems: multi-agent systems and geographic information systems. We also propose a generic model of agent-oriented simulation that we will apply to the field of land use planning. In fact, simulating the evolution of the urban system is a key to help decision makers to anticipate the needs of the city in terms of installing new equipment and opening new urbanization' areas to install the new population.

Keywords-Multi-agent system; Geographic information system; Modeling; Simulation; Complex system; Land use planning.

I. INTRODUCTION

A great part of the challenge of modeling and simulating interactions between natural and social processes has to do with the fact that processes in these systems result in complex temporal-spatial behavior. To deal with this problem, we propose a general architecture based on two modules:

- A GIS module: to instantiate the simulation engine.
- A MAS module: to represent the interactions between the different agents of the system

We also propose a generic agent-oriented model for simulating complex systems with spatial dynamics.

We apply our model to the field of land use planning. In fact, we will study the urban system and simulate its long term evolution through a backdrop of demographic change in a temporal-spatial heterogeneous environment.

Our objective is to understand how public policy can influence the transformation of cities, especially in regard to the opening of new urban' areas and installing new equipments.

II. STATE OF THE ART

In this section, we will detail the multi-agent systems, geographic information systems, complex systems and agent-based modeling. These points are the core of our design model simulation.

A. Multi agents approach

1) Notion of agent

An agent is a physical or virtual feature, which owns all or part of the following:[1]

- Located in an environment: means that the agent can receive sensory input from its environment and can perform actions that are likely to change this environment.
- Independent: means that the agent is able to act without the direct intervention of a human (or another agent) and he has control of its actions and its internal state.
- Flexible means that the agent is:
 - Able to respond in time: it can sense its environment and respond quickly to changes taking place.
 - Proactive: it does not simply act in response to its environment; it is also able to behave opportunistically, led by its aims or its utility function, and take initiatives when appropriate.
 - Social: it is capable of interacting with other agents (complete tasks or assist others to complete theirs)

2) Multi agents system

A multi-agents system (MAS) is a system composed of a set of agents situated in some environment and interacting according to some relations. There are four types of agent architecture [2]:

- Reactive agent: is responding to changes in the environment.
- Deliberative agent: makes some deliberation to choose its actions based on its goals.
- Hybrid agent: that includes a deliberative as well as a reactive component
- Learner agent: uses his perceptions not only to choose its actions, but also to improve its ability to act in the future.

B. Geographic Information System

1) Definition

According to the Environmental Systems Research Institute (ESRI), "a geographic information system (GIS) integrates hardware, software, and data for capturing, managing, analyzing, and displaying all forms of geographically referenced information."

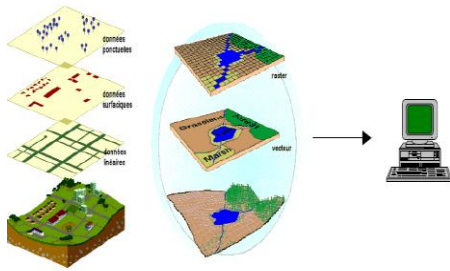


Figure 1. Geographic Information System.

2) Functions of GIS

The Functions of GIS describe the steps that have to be taken to implement a GIS. These steps have to be followed in order to obtain a systematic and efficient system. The steps involved are:

- **Data Capture:** Data used in GIS often come from many sources. Data sources are mainly obtained from Manual Digitization and Scanning of aerial photographs, paper maps, and existing digital data sets. Remote-sensing satellite imagery and GPS are promising data input sources for GIS.
- **Data Compilation:** Following the digitization of map features, the user completes the compilation phase by relating all spatial features to their respective attributes, and by cleaning up and correcting errors introduced as a result of the data conversion process.
- **Data Storage:** Once the data have been digitally compiled, digital map files in the GIS are stored on magnetic or other digital media. Data storage is based on a Generic Data Model that is used to convert map data into a digital form. The two most common types of data models are Raster and Vector.
- **Manipulation:** Once data are stored in a GIS, many manipulation options are available to users. These functions are often available in the form of "Toolkits."
- **Analysis:** The heart of GIS is the analytical capabilities of the system. The analysis functions use the spatial and non-spatial attributes in the database to answer questions about the real world. Geographic analysis facilitates the study of real-world processes by developing and applying models. Results of geographic analysis can be communicated with the help of maps, or both.

C. Complex system

A complex system is a set consisting of a large number of interacting entities that prevent the observer to predict its feedback, behavior or evolution by calculation. It is characterized by two main properties [3]-[4]:

- **Emergence** is the appearance of behavior that could not be anticipated from the knowledge of the parts of the system alone
- **Self-organization** means that there is no external controller or planner engineering the appearance of the emergent features

Traditional modeling approaches for complex systems focus on either temporal or spatial variation, but not both. To understand dynamic systems, patterns in time and space need to be examined together.

The proposed architecture for simulating complex systems that is used in this study, meshes these fields together in an approach called Simulating Complex System with Spatial Dynamics.

D. Agent-Based Modeling of Complex Systems with spatial dynamics

We note that the Agent-Based Modeling (ABM) is well suited to complex systems with spatial dynamics. In fact, the various components of these systems can be represented by agents.

We mention below four characteristics that show why Agent-Based Modeling is well suited to complex systems with spatial dynamics [4]-[5]:

- **Emergence:** ABMs allow to define the low-level behavior of each individual agent in order to let them interact (over time and space) to see whether some emergent property arises or not, and if it does, under which circumstances.
- **Self-Organization:** ABMs do not have any kind of central intelligence that governs all agents. On the contrary, the sole interaction among agents along with their feedbacks is what ultimately "controls" the system. This lack of a centralized control is what enables (and enforces) its self-organization.
- **Coupled Human-Natural Systems:** ABMs allow considering together both, social organizations with their human decision-making with biophysical processes and natural resources. This conjunction of subsystems enables ABMs to explore the interrelations between them, allowing analyzing the consequences of one over the other.
- **Spatially Explicit:** the feature of ABMs of being able to spatially represent an agent or a resource is of particular interest when communications and interactions among neighbors is a key issue. This can either imply some kind of internal representation of space or even the use of a Geographical Information System (GIS) with real data. This feature is of special interest in the case of agro-ecosystems.

III. AGENT-ORIENTED SIMULATION VS CLASSIC SIMULATION

A. Definitions

Simulate is to reproduce a phenomenon in order to [6]:

- Test hypotheses to explain the phenomenon
- Predicting the evolution of the phenomenon

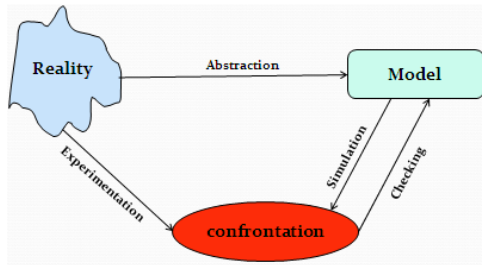


Figure 2. Simulation process.

Model is a simplified picture of reality that we use to understand the functioning of a system based on a question.

B. Classic simulation

1) Continuous models

Continuous time models are characterized by the fact that in a finite time interval, the system state variables change value continuously [6]-[7].

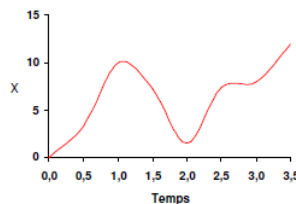


Figure 3. Continuous models

The simulation of continuous models encounters the difficulty of reproducing the continuity of the system dynamics due to the nature of the digital computer (solution: use of numerical integration methods)

2) Discrete models

The evolution of variables' state of the system is discretely at a time t after $t + dt$ [7].

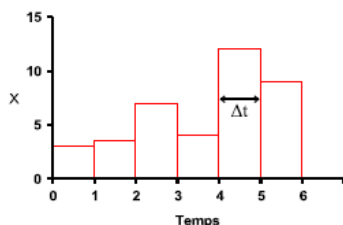


Figure 4. Discrete models

The simulation of discrete models can be summarized in two steps:

- Determine the function that implements the system dynamics.
- Increment time by one unit.
In this type of simulation the choice of the time step is very important because:
 - If a large time step:
Problem of management competitive events; Events in rapid occurrence are not considered
 - If small time step:
Problem of decomposition behavior in elementary behaviors

3) Discrete event models

The time axis is generally continuous; it's represented by a real number. However, unlike continuous models, the system state variables change discretely at precise moments that are called events.

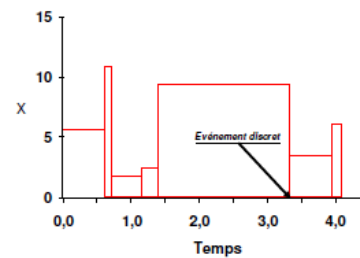


Figure 5. Discrete event models

In this type of simulation, there are three policies implementation:

- Scheduling of events: during the simulation, the future events that must occur on the system are predetermined;
- Analysis of activities: during the simulation, the events are not planned but triggered when certain conditions are met (collision of two vehicles, ...);
- Interaction process: this approach is the combination of the two others.

C. Limits of Classical Simulation

In the classical simulation, we found the following limits [6]:

- Equation in model with a large number of parameters;
- Difficulty of passing micro/macro, unable to represent different levels;
- No representation of behavior, but their results;
- Lack of realism (social sciences...);
- Does not explain the emergence of space-time structures;
- Difficulty of modeling the action.

D. Agent-Oriented Simulation

Agent-Oriented Simulation (AOS) is now used in a growing number of sectors, where it gradually replaces the various techniques of micro simulation and object-oriented simulation. This is due to its fundamental characteristics which are [6]-[9]:

- Representation as computer agents (state, skills, abilities,

resources);

- Representation of possible interactions between agents;
- Representation of space-time environment (agent is located).

This apparent versatility make the AOS the best choice for simulating complex systems and it spreads in a growing number of areas: sociology, biology, physics, chemistry, ecology, economics, land use planning, etc. The four aspects of a simulation model are:

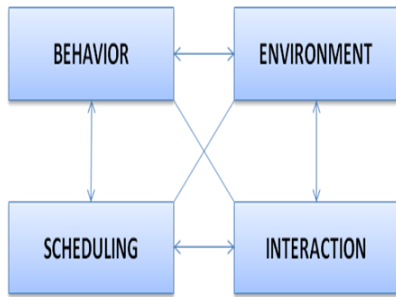


Figure 6. Agent oriented simulation model

- Module behavior: it relates to the deliberative process modeling agents;
- Module environment: the problem is to define the various physical objects in the world and the dynamics of the environment;
- Module scheduling: it concerns the modeling of the flow of time and the definition of scheduling used;
- Module interaction: it relates specifically to the modeling result of the actions and interactions they entail at time t.

IV. PROPOSED APPROACH

A. System Architecture

The simultaneous use of a geographical information system and a multi-agents system introduces an additional level of modeling complex system with spatial component. We propose the following architecture:

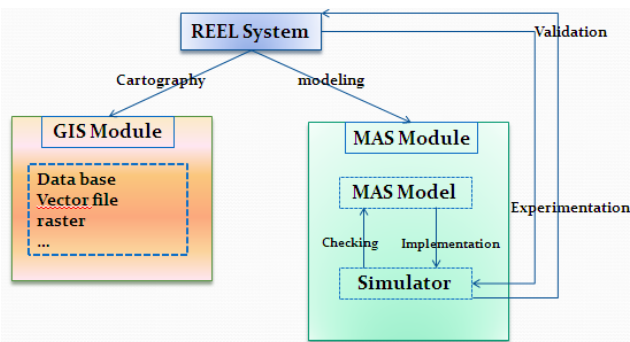


Figure 7. Proposed architecture.

- The module "GIS" is a descriptive representation of the reality that is used to initialize the multi-agent model, it contains the data to instantiate each one of the agents;
- The module "MAS" is used to model the real system and generate simulations involving agents that react and

interact and so it can test scenarios.

B. Generic model of agent-oriented simulation

To simulate complex systems with spatial component, we propose the following generic model:

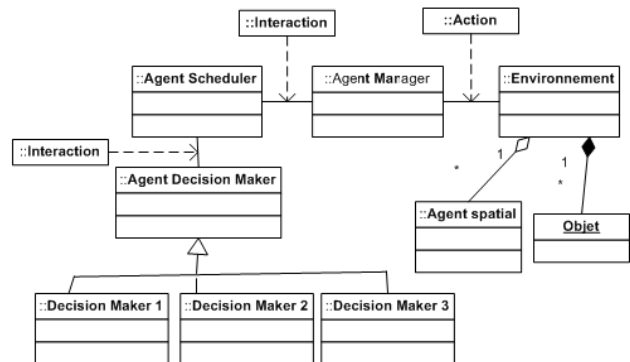


Figure 8. Generic model for agent-oriented simulation.

- The environment is heterogeneous composed of objects and spatial agents;
- Spatial Agent: This is the basic entity that represents the spatial part of our simulator.
- Agent manager: it is the agent that creates and acts on all environmental agents. Also, it interacts with the scheduler agent to determine the scheduling of its actions on the agents. It is responsible for the initialization of the simulation;
- Agent scheduler: Its role is to schedule the actions of the agents on the environment;
- Agent decision maker: it has several rules that must be applied to the environment

V. APPLICATION

A. Problematic of land use planning

We propose to apply our architecture to the field of land use planning. Our goal is to understand how public policy can influence the transformation of cities, especially in regard to the opening of new urban areas and installing new equipment. The simulation will determine when, during the dynamic of the city, public policy can act and have the most impact. As the modeling context we choose the city "Casablanca", and more specifically the district "Dar Bouazza".

1) Population dynamic

The population of Casablanca was established in the 2004 census, approximately to 3.63 million resident, 3 325 000 in urban areas and 305 000 in rural areas. It has an annual growth (50 500 persons per year)

For the district "Dar Bouazza", it has a population of 115367 residents with a growth rate of 1.8% [10].

2) Spatial dynamic

- Not enough housing: Housing remains a subject of major concern for the government. In fact, the insufficient offer of housing is a major problem in the field of land use planning.

- Lack of equipment in some region: the district “Dar bouazza” is experiencing an apparent lack of equipment compared to the normative grid of equipment to provide:

Nature of equipment	Users	Area (m ²)
Teatching		
primary school	1/8000	4000
school	1/16000	8000
high school	1/32000	10000
health		
Urban Health Center	1/30000	500-1000
Local Hospital	1/100000	15000-30000
Sports		
Sports area	1/45000	10000
Sports center	1/150000	50000
Public Services		
Marketplace	50000	2500 - 4000
Administratif		
Police station	1/45000	2000
civil Protection	1/100000	10000
Administrative Services	1/45000	3000
undefined	1/45000	3000
cultural		
mosque	1/15000	3500

TABLE I. NORMATIVE GRID OF EQUIPMENTS

- Lack of green spaces: Casablanca knows deficiencies of green spaces. The average is less than 1m² of public green space per resident, compared to the standard 10 m² per resident of the World Health Organization.

B. Agent-oriented simulation model

We apply our agent-oriented model in the context of land use planning, the system consists of: Agent Manager, Agent scheduler, LandUseAgent and two agents decision maker.

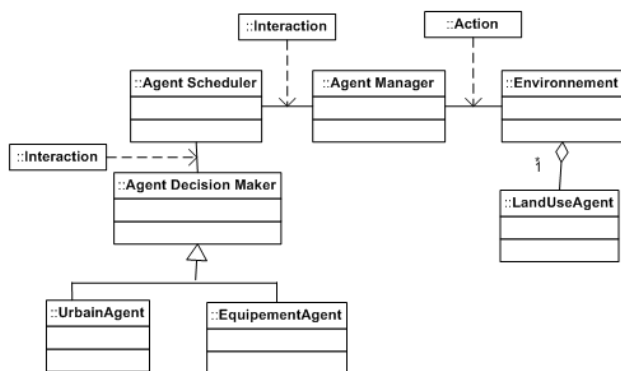


Figure 9. Agent-oriented model for simulation applied to the field of land use planning.

- Agent Manager: it initializes the simulation and creates displays. At each iteration, it executes the actions scheduled in the Agent Scheduler;

- Agent Scheduler: at each step, it called the actions of decision makers in a synchronous manner. The step in our model is one year;
- LandUseAgent: This is the entity that constitutes the spatial environment, it has:

An attribute "State": This contains information about the mode of land use. It can take the following values (bare land, land built, industrial area, green space)Attributes "potentiel_urban, potentiel equip, potentiel environnement": these attributes vary between 0-3 and indicate the potential of the cell to be intended to urbanization, to receive equipment or to be conserved as green space. The cell potential is calculated based on the master plan of Casablanca that indicates the destination overall soil over the next 30 years [10].

- UrbainAgent: its objective is to ensure urban sprawl for the installation of the new population; it obtained a surface demand. At each cycle, the agent performs the following operations:
 - Get the value of the new population;
 - Multiply this value by a stretch ratio set by the user;
 - Obtain a surface demand.

According to the constraints defined by the user agent changes the state of the cells by taking them from its target list (starting from the highest urban potential of the cells).

- EquipementAgent: its aim is to provide a number of facilities to suit the new population. At each cycle the agent performs the following operations:
 - Get the value of the new population;
 - Calculate the number of existing facilities;
 - Calculate the amount of equipment required for the installation of the new population;
 - Extract a demand.

According to the constraints defined by the user, EquipementAgent changes the state of the cells by taking them from its target list (starting from the highest potential of the cells).

A. Realization

To implement our model, we chose the multi-agent simulation platform Repast Symphony 2.0 [11] and the GIS software ArcGIS.

Our work environment is:

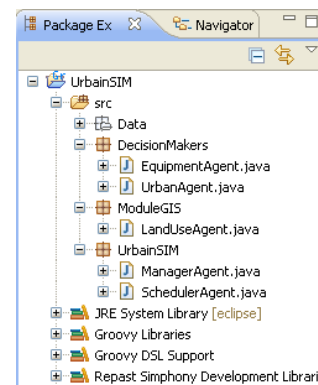


Figure 10. Environment work.

The user can specify the rate of urban growth and the ratio of urban sprawl by this interface:

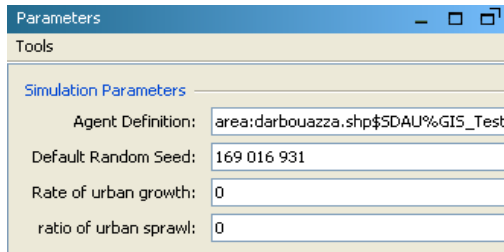


Figure 11.Environment work.

The initial state of the simulator is:

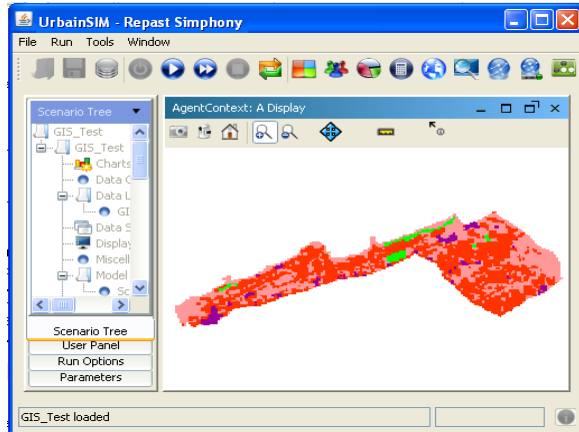


Figure 12.initial state of the simulator.

Legend :

- Bare land ■
- Built land ■
- Industrial area ■
- Green space ■

After twenty iterations, the population of “Dar Bouazza” has increased from 115 367 to 156 899 residents. To install the new population, UrbanAgent will open new areas to urbanization. It calculates the demand and changes a number of cells of the SpaceAgent (starting from the highest potential of the cells) to built land. The choice of cells to change is done in a random way. We obtain the following result:

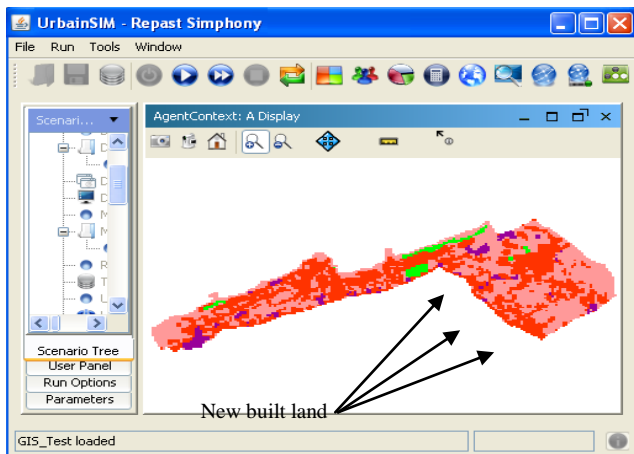


Figure 13.Results of the simulation after five iterations.

After 20 iterations, we see that the number of Equipment to create at "DarBouazza" is about twenty equipments as detailed in the diagram below:

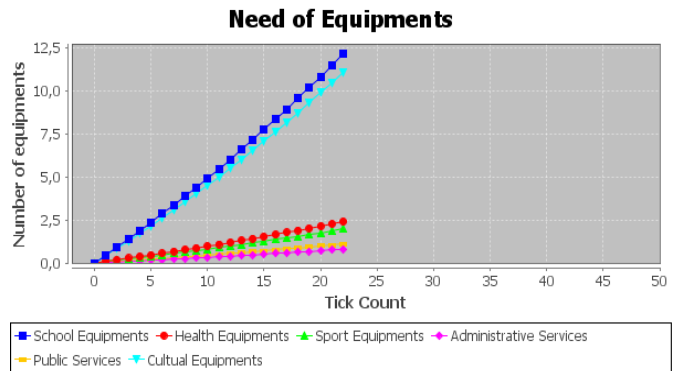


Figure 14. Need of equipments.

The installation of the new population will require the construction of:

- Eleven school equipments distributed as follows: Six primary schools, three schools, one high school.

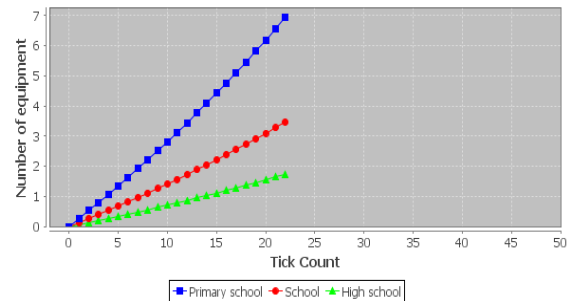


Figure 15. Need of school equipments.

- Ten cultualequipments (mosque)

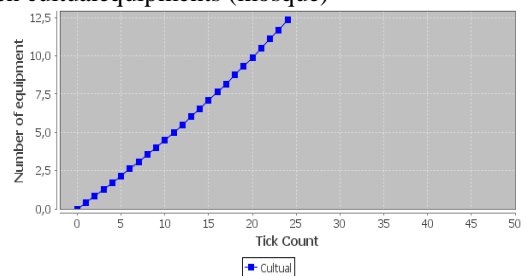


Figure 16. Need of cultualequipments.

- Two health equipments distributed as follows

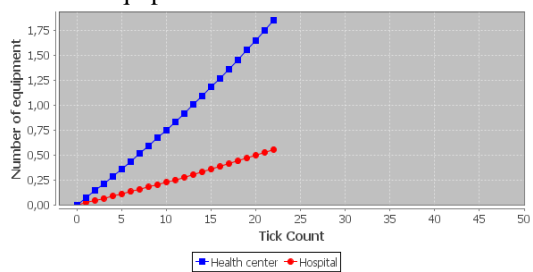


Figure 17. Need of health equipments.

- One sport equipment distributed as follows:

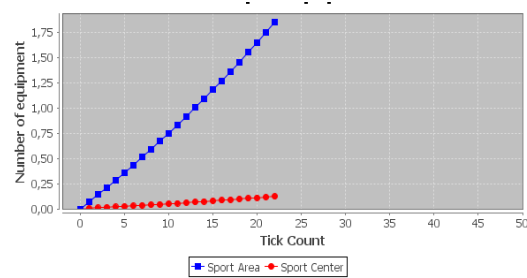


Figure 18. Need of sport equipments.

- Administrative services may not to be created as it's showed below:

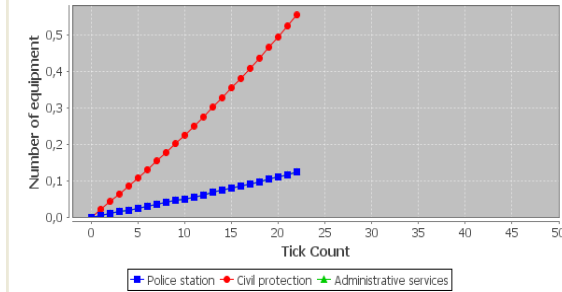


Figure 19. Need of administrative services.

- One public service:

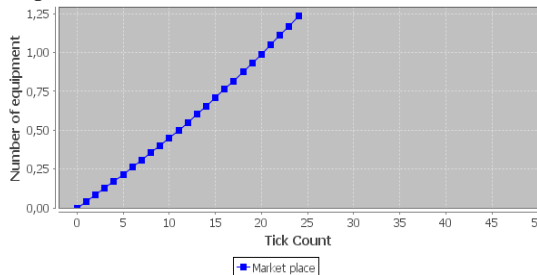


Figure 20. Need of public services.

VI. CONCLUSION

In this paper, we propose a new approach for simulating complex systems with spatial dynamics. First, we presented an architecture based on coupling between the two systems GIS and MAS. Then, we also proposed an agent-oriented simulation model. We applied our model to the field of land use planning. Our objective is to simulate the evolution of the city based on two dynamics: population dynamics and spatial dynamics. Results of the simulation show that after twenty iterations, it will be necessary to open about 40 ha to urbanization and to construct about twenty equipments.

As perspectives to our work, we will enrich the knowledge base of Decision maker agent.

In fact, more this model will be enriched by heating engineers (City planners, statisticians, sociologists ...) more it can help decision makers to have several simulation scenarios. Also, we can detail the agent scheduler to show how the scheduling system can influence the outcome of the simulation.

REFERENCES

- [1] B. Chaib-Ddraa, I. Jarras, B. Moulin, Systèmes multiagents : Principes généraux et applications , (2001) Hermès.
- [2] J. Ferber, "Les systèmes multi-agents, vers une intelligence collective", Inter Editions (1995).
- [3] Ing. Jorge Corral Areán, "agent-based methodology for developing agroecosystems' simulations", university of republica (2011).
- [4] CSIRO, "Complex or just complicated: what is a complex system?": csiro.au [Online]. Available: www.csiro.au/resources/AboutComplexSystems.html[Accessed: Sept. , 2011].
- [5] T. Moncion, "Modélisation de la complexité et de la dynamique des simulations multi-agents. Application pour l'analyse des phénomènes émergents", PhDthesis, Université d'Evry val d'Essonne, (2008).
- [6] J. Ferber, "Agent-based simulation", LIRMM - Université Montpellier II, (2009).
- [7] D. Davis, "Prospective Territoriale par Simulation Orientée Agent", PhDthesis, La Reunion university, (2011).
- [8] P. Dawn, A. Hessel, S. Davisl, "Complexity, land-use modeling, and the human dimension: Fundamental challenges for mapping unknown outcome spaces," Geoforum, Volume 39 Issue 2: 789-804., 2008.
- [9] F. Bousquet, C. Le Page, "Multi-agent simulations and ecosystem management: a review", Elsevier, (2004).
- [10] IAURIF, "Plan de développement stratégique et schéma directeur de Casablanca", AUC (2008).
- [11] Repast documentation, repast.sourceforge.net [Online]. Available: http://repast.sourceforge.net/docs.html[Accessed: October. , 2012].

Fatimazahra, Barramou received her engineer degree in Geographic Information System in 2009 at the Hassania School of Public Works (EHTP :EcoleHassania des Travaux Publics), Casablanca, Morocco. In 2010 she joined the system architecture team of the National and High School of Electricity and Mechanic (ENSEM: Ecole Nationale Supérieure d' Electricité et de Mécanique), Casablanca, Morocco. Her actual main research interests concern Modeling and simulating complex systems based on Multi-Agent Systems and GIS.

Ms. Barramou is actually a Software Engineer in the Urban Agency of Casablanca.

Malika, Addou received her Ph.D. in Artificial Intelligence from University of Liege, Liege, Belgium, in 1992. She got her engineer degree in Computer Systems from the Mohammadia School of Engineers (EMI :Ecole Mohammadia des ingénieurs), Rabat, Morocco in 1982. She is Professor of Computer Science at the Hassania School of Public Works (EHTP :EcoleHassania des Travaux Publics), Casablanca, since 1982. Her research focuses on Software Engineering (methods and technologies for design and development), on Information Systems (Distributed Systems) and on Artificial Intelligence (especially Multi-Agent Systems technologies).

Improving Web Page Prediction Using Default Rule Selection

Thanakorn Pamutha, Chom Kimpan
Faculty of Information Technology
Rangsit University, RSU
Phatumthani, THAILAND

Siriporn Chimplee
Faculty of Science and Technology
Suan Dusit Rajabhat University, SDU
Bangkok, THAILAND

Parinya Sanguansat
Faculty of Engineering and Technology
Panyapiwat Institute of Management, PIM
Nonthaburi, THAILAND

Abstract—Mining user patterns of web log files can provide significant and useful informative knowledge. A large amount of research has been done in trying to predict correctly the pages a user will most likely request next. Markov models are the most commonly used approaches for this type of web access prediction. Web page prediction requires the development of models that can predict a user's next access to a web server. Many researchers have proposed a novel approach that integrates Markov models, association rules and clustering in web site access predictability. The low order Markov models provide higher coverage, but these are couched in ambiguous rules. In this paper, we introduce the use of default rule in resolving web access ambiguous predictions. This method could provide better prediction than using the individual traditional models. The results have shown that the default rule increases the accuracy and model-accuracy of web page access predictions. It also applies to association rules and the other combined models.

Keywords—web mining, web usage mining; user navigation session; Markov model; association rules; Web page prediction; rule-selection methods.

I. INTRODUCTION

The rapid expansion of the World Wide Web has created an unprecedented opportunity to disseminate and gather information online. There is an increasing need to study web-user behavior to improve the service to web users and increase their value of enterprise. One important data source for this study is the web-server log data that traces the user's web browsing actions [1]. Predicting web-users' behavior and their next movement has been recognized and discussed by many researchers lately. The need to predict the next Web page to be accessed by the users is apparent in most web applications today whether or not they are utilized as search engines, e-commerce solutions or mere marketing sites. Web applications today are designed to provide a more personalized experience for their users [2]. The result of accurate predictions can be used for recommending products to customers, suggesting useful links, as well as pre-sending, pre-fetching and caching of web pages in reducing access latency [1]. There are various ways that can help us make such a prediction, but the most common approaches are the Markov models and association

rules. Markov models are used in identifying the next page to be accessed by the Web site user based on the sequence of their previously accessed pages. Association rules can be used to decide the next likely web page requests based on significant statistical correlations.

Yang, Li and Wang [1] have studied different association-rule based methods for web request predictions. In this work, they have examined two important dimensions in building prediction models, namely, the type of antecedents of rules and the criteria for selecting prediction rules. In one dimension, they have a spectrum of rule representation methods which are: subset rules, subsequence rules, latest subsequence rules, substring rules and latest substring rules. In the second dimension, they have rule-selection methods namely: longest-match, most-confident and pessimistic selection. They have concluded that the latest substring representation, coupled with the pessimistic-selection method, gives the best prediction performance. The authors [2-5] have applied the latest substring representation using the most-confident selection method to building association-rule based prediction models from web-log data using association rules. In Markov models, the transition probability matrix is built making and predictions for web sessions are straight forward. In Markov models, the target is to build prediction models to predict web pages that may be next requested, The consequence to this is that, only the highest condition probability is considered. Hence, a prediction rule with the highest condition probability is chosen.

In the past, researchers have proposed different methods in building association-rule based prediction models using web logs, but none had yielded significant results. In this paper, we propose the default rule in resolving ambiguous predictions. This method could provide better prediction than using the traditional models individually.

The rest of this paper is organized as follows:

- Sec.2 Related Works
- Sec.3 Markov Models
- Sec.4 Ambiguous rules
- Sec.5 Experimental Setup and Result
- Sec. 6 Conclusion

II. RELATED WORK

There are wide application areas in analyzing the user web navigation behaviors in web usage mining [6]. The analysis of user web navigation behavior can help improve the organization of web sites and web performance by pre-fetching and caching the most probable next web page access. Web Personalization and Adaptive web sites are some of the applications often used in web usage mining. Web usage mining can provide guidelines in improving e-commerce to handle business specific issues like customer satisfaction, brand loyalty and marketing awareness.

The most widely used approach is web usage mining that includes many models like the Markov models, association rules and clustering [5]. However, there are some challenges with the current state of the art solutions when it comes to accuracy, coverage and performance. A Markov model is a popular approach to predict what pages are likely to be accessed next. Many authors have proposed models for modeling the user web navigation sessions. Yang, Li and Wang [1] have studied five different representations of Association rules which are: subset rules, subsequent rules, latest subsequence rules, substring rules and latest substring rules. As a result of the experiments, performed by the authors concerning the precision of these five association rules representations using different selection methods, the latest substring rules were proven to have the highest precision. Deshpande and Karypis [7] have developed techniques for intelligently combining different order Markov models resulting to lower state complexity, improved prediction accuracy, while retaining the coverage of the All-Kth-Order Markov model. Three approaches had been widely used namely frequency-pruning, error-pruning and support-pruning to reduce the state space complexity. Khalil, Li and Wang [2] have proposed a new framework in predicting the next web page access. They used lower all k-th Markov models to predict the next page to be accessed. If the Markov model is not able to predict the next page access, then the association rules are used to predict the next web page. They have also proposed, on the other hand, solutions for prediction ambiguities. Chimphee [5] has proposed a hybrid prediction model (HyMFM) that integrates Markov model, Association rules and Fuzzy Adaptive Resonance Theory (Fuzzy ART) clustering all together. These three approaches are integrated to maximize their strengths. This model could provide better prediction than using an individual approach. In fact, Khalil, Li and Wang [4] had introduced the Integration Prediction Model (IPM) by combining the Markov model, association rules and clustering algorithm together. The prediction is performed on the cluster sets rather than the actual sessions. A. Anitha [8] has proposed to integrate a Markov model based sequential pattern mining with clustering. A variant of Markov model called the dynamic support, pruned all kth order in order to reduce state space complexity. The proposed model provides accurate recommendations with reduced state space complexity. Mayil [9] has proposed to model users' navigation records as inferred from log data, A Markov model and an algorithm scans the model first in order to find the higher probability trails which correspond to the users' preferred web navigation trails.

Predicting a user's next access on a website has attracted a lot of research work lately due to the positive impact of predictions on the different areas of web based applications [4]. First, many of the papers proposed using association rules or Markov models for next page predictions [1, 7, 9-11]. Second, many papers have addressed the uses of combining both methodologies [2-5]. Third, many of the papers have addressed the integration Markov model, Association rules and clustering method. This model could provide better predictions than using each approach individually [5]. It can be inferred that most researchers use Markov models for this type of prediction and is thus, mostly preferred.

In this paper, we study Markov models in predicting a user's next web request using default rule. The prediction models are based on web log data that correspond with users' behavior. They are used to make predictions for the general user more fit for a particular client. This prediction requires the discovery of a web users' sequential access patterns and using these patterns to make predictions of users' future access more accurate. We can then incorporate these predictions into web pre-fetching system in an attempt to enhance the performance.

III. MARKOV MODELS

Markov models are commonly used method for modeling scholastic sequences with underlying finite-state structures that are shown to be well-suited for modeling and predicting a user's browsing behavior on a web site [7]. The identification of the next web page to be accessed by a user is calculated based on the sequence of previously accessed pages.

In general, the input for this problem is the sequence of web pages that were accessed by a user. It is assumed that it discusses the Markov property. In such a process, the past is irrelevant in predicting the future given the knowledge of the present. Let $P = \{p_1, p_2, \dots, p_m\}$ be a set of pages in a web site. Let W be a user session including a sequence of pages visited by the user in a visit. Assuming that the user has visited l pages then $P(p_i/W)$ is the probability that the user visits page p_i next. Thus an equation may thus be deduced:

$$p_{l+1} = \operatorname{argmax}_{p \in P} \{P(P_{l+1} = p/W)\} \\ = \operatorname{argmax}_{p \in P} \{P(P_{l+1} = p/P_l, P_{l-1}, \dots, P_1)\} \quad (1)$$

Essentially, this approach for each symbol P_i computes its probability of being accessed next which then selects the web page that has the highest probability of accessibility. The probability, $P(p_i/W)$, is estimated by using all W . Naturally, the longer l and the larger W are, the more accurate the results are $P(p_i/W)$. However, it is not feasible to accurately determine these conditional probabilities because the sequences may arbitrarily be longer (or longer l), and the size of the training set is often much smaller than that required to accurately estimate the various conditional probabilities for long sequences (or large W). For this reason, the various conditional probabilities are commonly estimated by assuming that the process generating sequences of the web pages visited by users follows a Markov process. That is, the probability of visiting a web page p_i does not depend on all the pages in the web session, but only on a small set of k preceding pages,

where $k \ll l$. Using the Markov process assumption, the web page p_{l+1} will be generated next is given by

$$p_{l+1} = \operatorname{argmax}_{p \in P} \{P(P_{l+1} = p | P_l, P_{l-1}, \dots, P_{l-(k-1)})\} \quad (2)$$

Where k denotes the number of the preceding pages as it identifies the order of Markov model. The resulting model of this equation is called the k^{th} -order Markov model. In order to use the k^{th} -order Markov model the learning of P_{l+1} is needed for each sequence of k web pages.

Let S_j^k be a state with k as the number of preceding pages denoting the Markov model order and j as the number of unique pages on a web site, $S_j^k = \langle p_{l-(k-1)}, P_{l-(k-2)}, \dots, P_l \rangle$, by estimating the various conditional probability $P(P_{l+1} = p | P_l, P_{l-1}, \dots, P_{l-(k-1)})$. Using the maximum likelihood principle [5], the conditional probability $P(p_i | S_j^k)$ is computed by counting the number of times a sequence S_j^k occurs in the training set, and the number of times p_i occurs immediately after S_j^k . The conditional probability in the ratio of these two frequencies, therefore,

$$P(p_i | S_j^k) = \frac{\text{Frequency}(\langle S_j^k, p_i \rangle)}{\text{Frequency}(S_j^k)} \quad (3)$$

IV. AMBIGUOUS RULES

The main problem in association rules in the application of large data item sets is the discovery of a large number of rules and the difficulty in identifying the area that leads to the correct prediction. As regards the non-Markov models, they lack web page prediction accuracy because they do not use enough history in web access whereas Markov models have a high state space complexity [2-5].

There is an apparent direct relationship between Markov models and association rules techniques. According to the Markov model pruning methods presented by Deshpande and Karypis [7] and the association rules selection methods presented by Yang, Li and Wang [1].

The researchers herein have proposed different methods for building association-rule based prediction models in using web logs, but ambiguous rules still exist. In order to solve this problem, we propose the use of default rule to keep both the low state complexity and high accuracy results. We use the following examples to show the idea creating default rules. Consider the set of user session in table 1. Note that the numbers are assigned to web page names. Table 1 examines the following 6 user session:

TABLE I. USER SESSIONS

S1	900, 586, 594, 618
S2	900, 868, 594
S3	868, 586, 594, 618
S4	594, 619, 618
S5	868, 594, 900, 618
S6	868, 586, 618, 594, 619

Table 2 shows an example of counting the support of extracted web access sequences and the useful sequences are highlighted. The row represents the previously visited page

and the column represents the next visited page. Each field in the matrix is produced by looking at the number of times the web page on the horizontal line followed by the web page on the vertical. In this example, web user's session, web page 586 and web page 594 co-occurred in session S1 and S3, and therefore web page (A,B) has the support of 2.

TABLE II. AN EXAMPLE OF EXTRACTED WEB ACCESS SEQUENCES AND THEIR SUPPORT COUNT OF FIRST-ORDER MARKOV MODELS

Second item in sequence	Support count				
First item in sequence	586	594	618	619	868
586		2	1		
594			2	2	
618		1			
619			1		
868	2	2			
900	1		1		1

The next step is to generate rules from these remaining sequences. All the remaining sequences construct the prediction rules using the maximum likelihood principle. The condition probabilities of all sequences are calculated and are ranked.

For example, given the antecedent that the web page is 586, the condition probability is that 586 -> 594 is 85.7%. This is calculated by dividing the support value of the sequence (586,594) with the support value of the web page (586); (2/3 = 66.7%). From the training if antecedent web page is 586, then the single-consequence can be 594 and 618 with their confidence levels at 66.7%) and 33.3%. In this case, 586->594 have the highest probability value. Then a prediction rule with the highest condition probability is chosen.

Applying the First-order Markov models to the above training user sessions, we notice that the most frequent state is <594> and it appeared 6 times as follows:

$$P_{l+1} = \operatorname{argmax}\{P(618|594)\} = 618 \text{ OR } 619$$

Using Markov models, we can determine that there is a 50/50 chance that the next page to be accessed by the user after accessing page 594 could be either 618 or 619.

Obviously, this information alone does not provide us with correct predictions on the next page to be accessed by the user as we may have obtained the highest condition probability for both pages, using the 618 and 619.

To break the tie and find out which page would lead to the most accurate prediction, we choose the rule whose right hand side (RHS) is the most popular page in the training web log (page 618 = 5, page 619 = 2). Thus, we choose rule 594->618. We refer to this rule as the default rule.

In this paper, we introduce the default rule in resolving ambiguous predictions. It can apply to all Markov models and Association rule. This method avoids the complexity of high order Markov model. This method also improves the efficiency of web access predictions. Algorithm is summarized as follows:

```
Training:
Generate rule using the First-order Markov model
Test:
FOR each session of Test set
  Latest substring from session (got LHS->RHS)
  Compare with appropriate rules
  IF the matching rule provides an non-
ambiguous prediction
    THEN
      The prediction is made by the state
    ELSE //The ambiguous occurs
      Select rule whose RHS is the most
frequency the training web log THEN make
prediction
    ENDIF
  ENDFOR
```

In this work, we define an ambiguous prediction as one predictive page. This task could apply for an ambiguous prediction as two or more predictive pages that have the same conditional probabilities by a Markov model. The ambiguous prediction potentially has other definitions, for example, the certainty of a prediction is below a threshold. We, nonetheless, did not explore the other options in this paper [2].

V. EXPERIMENTAL SETUP AND RESULT

The experiment used on web data, as collected from the web server at the NASA Kennedy space Centre from 00:00:00 Aug 1, 1995 through 23:59:59 Aug 31, 1995, yielded a total of 31 days. During this period there are totally 1,569,898 requests recorded by the log file (see example in Fig.1).

Before doing the experiments, we removed all the dynamically generated content such as CGI scripts. We also filtered out requests with unsuccessful HTTP response codes [1, 12]. See Web log pre-processing in [12].

```
199.72.81.55 - - [01/Jul/1995:00:00:01 -0400] "GET
/history/apollo/ HTTP/1.0" 200 6245
```

Figure 1. A fragment from the server logs file

The next step that we had used was to identify the user sessions. The session identification splits all the pages accessed by the IP address which is a unique identity and a timeout, where if the time between page requests exceeds a certain limit, it is assumed that the user has started a new session.

A 30 minute default timeout is considered. Figure 2 shows a fragment from this session identification result. There were a total of 71,242 unique visiting IP addresses, 935 unique pages and 130,976 sessions. Also, the frequency of each page visited by the user was calculated. The page access frequency is shown in Figure 3 which reveals that page number 295 is the most frequent page accessed at and it was accessed 41109 times.

```
S1: 634, 391, 396, 408
S2: 393, 392, 400, 398, 396, 408, 37, 53
S3: 91, 124, 206, 101, 42, 287, 277
S4: 634, 391, 396, 631
S5: 124, 125, 127, 123, 130, 126, 131, 128, 129, 83
S6: 391, 634, 633, 295
S7: 295, 277, 91, 919
S8: 935, 391, 631, 634
S9: 755, 295, 810
S10: 637, 391, 918
```

Figure 2. A fragment from session identification result

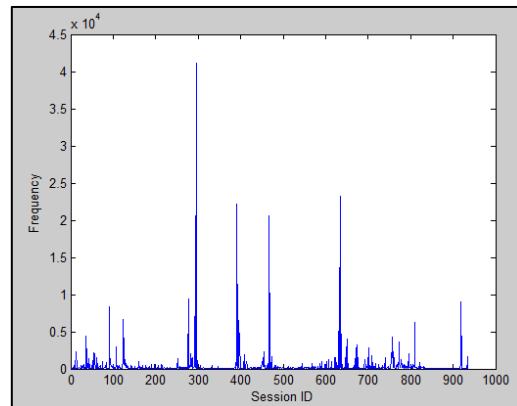


Figure 3. A frequency chart for the frequency visited sessions

All experiments were conducted on P i-7 2.20 GHz PC with 8 GB DDR RAM, running Windows 7 Home Premium. The algorithms were implemented using MATLAB.

This model will then be evaluated for accuracy on a previously unseen set of web sessions. These are called test set. When applying the trained model on the testing sequence, this is done by hiding the last symbol in each of the test sessions, and using the model to make a prediction of this trimmed session. During the testing step the model is given a trimmed sequence for prediction in which the last symbol of the sequence to compute the accuracy of the model is made. If the prediction model was unable to make a prediction for a particular web session, it was calculated as the wrong prediction. In the experiment, the proposed measure is compared with prediction model based evaluation that measures accuracy, coverage, and model accuracy [5]. Accuracy is defined as the ratio of the number of sequence for which the model is able to correctly predict the hidden symbol to the total number of sequence in the test set. The second is the coverage. It is defined as the ratio of the number of sequences whose required number for making a prediction was found in the model to the total number of sequences in the test set. The Third one is the model accuracy; it is calculated only on the web user sessions upon which the prediction was made. If the prediction model was unable to make a prediction for a particular web session, it is ignored in the model accuracy calculations.

The evaluation was calculated using a 10-fold cross validation. The data was split into ten equal sets. The first nine sets are considered as training data for rule constructions. Then, the second last set is considered as testing data for evaluation. The test set is continued moving upward until the first set was used for testing and the rest for training. The reported accuracy is the average of ten sets. The experiment results are the average of ten tests. We experimentally evaluated the performance of the proposed approach first-order Markov model and construct the predictive model.

Figure 4 shows the number of ambiguous rules (%). It shows that as the minimum support threshold has sequences varied (2-8), the number of ambiguous rules occurs at any support threshold.

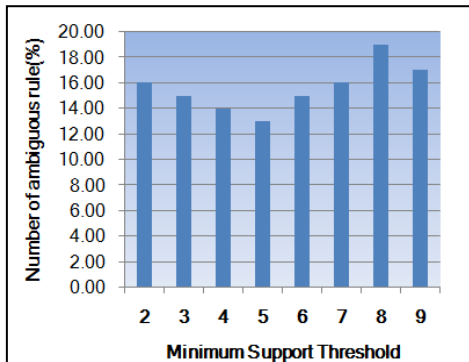


Figure 4. Number of ambiguous rules (%) occurred at difference minimum support thresholds.

The results are plotted in Figure 5-7. It shows that as the support threshold has increased, the coverage of the model decreased as accompanied by ad decrease in the accuracy. However, the model-accuracy of the model continues to increase.

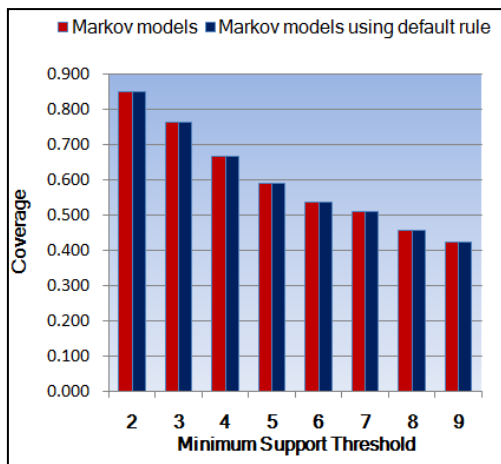


Figure 5. The comparison of coverage achieved by the Markov models and Markov models using default rules with difference minimum thresholds.

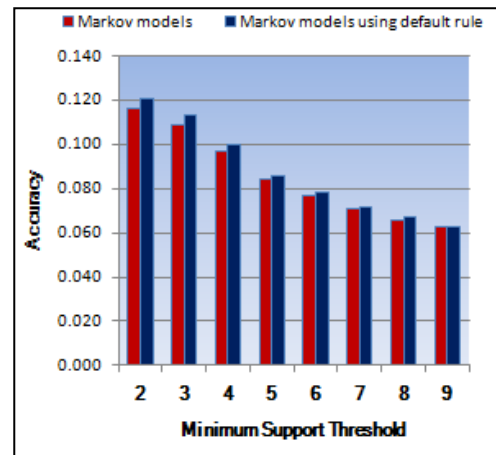


Figure 6. The comparison of accuracy achieved by the Markov models and Markov models using default rules with difference minimum support threshold.

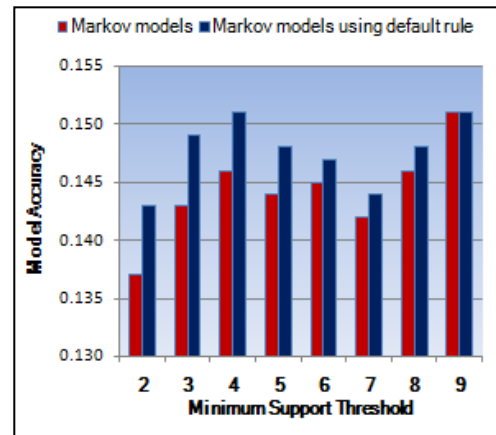


Figure 7. The comparison of Model-accuracy achieved by the Markov models and Markov models using default rule with difference minimum support threshold.

As it can be seen from Figures 5 – 7, the interesting observations can thus be summarized as the overall performances of Markov models using the default rule accuracy and model-accuracy while keeping their higher coverage abilities.

VI. CONCLUSION

A Markov models is a popular approach to predict what web page are likely to be accessed next. Many researchers have proposed to use the Markov models in predicting a web user's navigation session but there has been no proposal at present on how to solve the ambiguous problem rules. In this paper, we propose to use the default rules in resolving ambiguous prediction in the first order Markov models. This method could provide better web navigation prediction than merely using the individual traditional models individually. It can also apply to all Markov models, Association rule and the other combined Markov models.

ACKNOWLEDGMENT

The authors gratefully thank the anonymous referees and collaborators for their substantive suggestions. We also acknowledge research support from Rangsit University at Bangkok, Thailand.

REFERENCES

- [1] Q. Yang, T. Li and K. Wang, "Building Association -Rules Based Sequential Classifiers for Web-Document Prediction," *Journal of Data Mining and Knowledge Discovery*, Vol. 8, 2004.
- [2] F. Khalil, J. Li and H. Wang, "A Framework of Combining Markov Model with Association Rules for Predicting Web Page Accesses," *Conferences in Research and Practice in Information Technology (CRPIT)*, Vol. 61, 2006.
- [3] S. Chimphee, N. Salim, M.S.B Ngadiman, W. Chimphee and S. Srinoy, "Predicting Next Page Access by Markov Models and Association Rules on Web Log Data," *The international Journal of Applied Management and Technology*, Vol. 4, 2006.
- [4] F. Khalil, J. Li and H. Wang, "Integrating recommendation models for improved web page prediction accuracy," *Thirty-First Australasian Computer Science Conference (ACSC2008)*, 2008.
- [5] S. Chimphee, N. Salim, M.S.B. Ngadiman and W.Chimphee, "Hybrid Web Page Prediction Model for Predicting a User's Next Access," *Information Technology Journal*, Vol.9, No. 4, 2010.
- [6] Bhawna Nigam, S.J.a.S.T., "Mining Association Rules from Web Logs by Incorporating Structural Knowledge of Website," *International Journal of Computer Applications (0975-8887)*, 2012. Vol. 42(No. 11): p. pp. 17-23.
- [7] M. Deshpande and G. Karypis, "Selective Markov Models for Predicting Web Page Accesses," *In ACM Transactions on Internet Technology*, Vol.4, No.2,2004.
- [8] A. Anitha, "A New Web Usage Mining Approach for Next Page Access Prediction," *International Journal of Computer Applications (0975-8887)*, Vol. 8, No. 11, 2010.
- [9] V.V. Mayil, "Web Navigation Path Pattern Prediction using First Order Markov Model and Depth first Evaluation," *International Journal of Computer Applications(0975-8887)*, Vol. 45, No. 16, 2012.
- [10] S. Venkateswari and R.M. Suresh, "Association Rule Mining in E-commerce: A Survey," *International Journal of Engineering Science and Technology (IJEST)*, Vol. 3, No. 4, 2011.
- [11] N.K. Tyagi and A.K. Solanki, "Prediction of Users Behavior through Correlation Rules," *International of Advanced Computer Science and Applications (IJACSA)*, Vol. 2, No. 9, 2011.

- [12] T. Pamutha, S. Chimphee, Ch. Kimpan and P. Sanguansat, "Data Preprocessing on Web Server Log Files for Mining Users Access Patterns," *International Journal of Research and Reviews in Wireless Communications (IJRRWC)*, Vol.2, No.2, 2012.

AUTHORS PROFILE



Thanakorn Pamutha received the M.Sc. degree in Computer Science from Prince of Songkla University, Songkla-Thailand in 2000. He has been an assistant professor in Yala Rajabhat University, Thailand since 2003. Now, he is a PhD student in Information Technology, Faculty of Information Technology, Rangsit University, Thailand. He is interested in database system, artificial intelligence, data mining, and web usage mining.



Siriporn Chimplee received the M.Sc. degree in Applied Statistics from National Institution Development of Administration (NIDA), Thailand, and Ph.D. degrees in Computer Science from University Technology Malaysia, Malaysia. She is a lecturer at the Computer Science Department, Faculty of Science and Technology, Suan Dusit Rajabhat University, Thailand. She is interested in web mining, web usage mining, statistical, and soft computing.

data mining, web usage mining, statistical, and soft computing.



Chom Kimpan received his D.Eng in Electrical and Computer Engineering from King Mongkut's Institute of Technology Ladkrabang, M.Sc in Electrical Engineering from Nihon University, Japan, and bachelor's degree in Electrical Engineering from King Mongkut's Institute of Technology of Thailand. Now he is an Associate Professor at the Department of Informatics, Faculty of

Information Technology, Rangsit University, Thailand. He is interested in pattern recognition, image retrieval, speech recognition, and swarm intelligence.



Parinya Sanguansat received the B.Eng., M.Eng. and Ph.D. degrees in Electrical Engineering from the Chulalongkorn University, Thailand. He is an assistant professor in the Faculty of Engineering and Technology, Panyapiwat Institute of Management, Thailand in 2001, 2004 and 2007 respectively. His research areas are digital signal processing in pattern recognition including on-line handwritten recognition, face and automatic target recognition.

on-line handwritten recognition, face and automatic target recognition.

Multilayer Neural Networks and Nearest Neighbor Classifier Performances for Image Annotation

Mustapha OUJAOURA

Laboratory of Information
Processing and Telecommunications,
Computer Science Department,
Faculty of Science and Technology,
Sultan Moulay Slimane University,
PO Box. 523, Béni Mellal, Morocco.

Brahim MINAOUI

Laboratory of Information
Processing and Telecommunications,
Computer Science Department,
Faculty of Science and Technology,
Sultan Moulay Slimane University,
PO Box. 523, Béni Mellal, Morocco.

Mohammed FAKIR

Laboratory of Information
Processing and Telecommunications,
Computer Science Department,
Faculty of Science and Technology,
Sultan Moulay Slimane University,
PO Box. 523, Béni Mellal, Morocco.

Abstract—The explosive growth of image data leads to the research and development of image content searching and indexing systems. Image annotation systems aim at annotating automatically an image with some controlled keywords that can be used for indexing and retrieval of images. This paper presents a comparative evaluation of the image content annotation system by using the multilayer neural networks and the nearest neighbour classifier. The region growing segmentation is used to separate objects, the Hu moments, Legendre moments and Zernike moments which are used in as feature descriptors for the image content characterization and annotation. The ETH-80 database image is used in the experiments here. The best annotation rate is achieved by using Legendre moments as feature extraction method and the multilayer neural network as a classifier.

Keywords—Image annotation; region growing segmentation; multilayer neural network classifier; nearest neighbour classifier; Zernike moments; Legendre moments; Hu moments; ETH-80 database.

I. INTRODUCTION

As the online resources are still a vital resource in everyday life, developing automatic methods for managing large volumes of digital information is increasingly important. Among these methods, automatic indexing multimedia data remain an important challenge, especially the image annotation [1]. Annotated images play a very important role in information processing. They are useful for an image retrieval based on keywords and image content [2]. Manual annotation is not only boring but also not practical in many cases due to the abundance of information. Most images are, therefore, available without adequate annotation. Automatic image content annotation becomes a recent research interest [3]. It attempts to explore the visual characteristics of images and associate them with image contents and semantics.

Many classifiers have been used for the image classification and annotation without any performance's evaluation. In this paper, we use the neural network classifier for image annotation, and we compare its performance with the nearest neighbour classifier. We use the same type of moments as a method of features extraction for each classifier in order to have an objective comparison between

performances of the considered classifiers. In such case, we used a system that tends to extract objects from each image and find the annotation terms that describe its individual content.

The rest of the paper is organised as follows. The Section 2 presents the adopted annotation system while the Section 3 discusses the primordial tasks of any annotation and recognition system. They are the image segmentation and features extraction problems in addition to a brief formulation for Hu, Legendre and Zernike moments as features extraction method. The Section 4 is reserved for the image classification and annotation using neural network or the nearest neighbour classifier. Finally, the Section 5 presents the experimental annotation results and the comparing performance of each used classifier.

II. ANNOTATION SYSTEM

The automatic annotation's techniques of image's content attempt to explore visual features of images that describe an image content and associate them with its semantics. It's an effective technology to improve the image indexing and searching in the large volume of information that is available in the media. The algorithms and systems, used for image annotation, are commonly divided into those tasks [4]:

- Segmentation and Features extraction;
- Classification and Annotation.

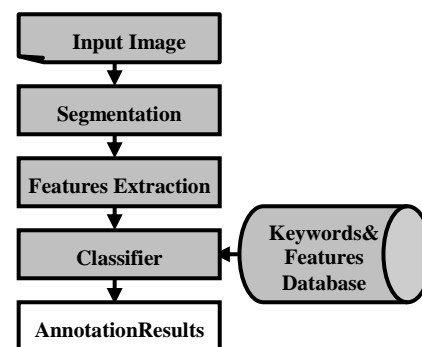


Figure 1. The block diagram of the image annotation system.

The annotation system adopted in this work is shown in Fig.1. The system has a reference database that contains keywords and features descriptor of images that are already annotated by experts (manual offline annotation). This database is used for modelling and training the classifier in order to choose the appropriate keywords. To achieve this goal, by using region growing segmentation, the input image is firstly segmented into regions that represent objects in the image. The feature vectors of each region are secondly computed and extracted from the image. Those features are finally feed into the input of the classifier, that can be the multilayer neural networks or the nearest neighbor classifier, in order to decide and choose the appropriated keywords for annotation tasks. Therefore, the image content is annotated and the performance of each classifier can be evaluated.

III. IMAGE SEGMENTATION AND FEATURES EXTRACTION

The features' vector, which is extracted from the entire image, loses local information. So, it is necessary to segment an image into regions or objects of interest and use of local characteristics. The image segmentation is the process of partitioning a digital image into multiple segments. It is very important in many applications for any image processing, and it still remains a challenge for scientists and researchers. So far, the efforts and attempts are still being made to improve the segmentation techniques. With the improvement of computer processing capabilities, there are several possible segmentation techniques of an image: threshold, region growing, active contours, level sets, etc... [5]. Among these methods, the region growing is well suited because of its simplicity and has been successfully used several times as a segmentation technique of digital images.

The regions are iteratively grown by comparing all unallocated neighbouring pixels to the regions. The difference between a pixel's intensity and the pixel's mean in one region is used as a measure of similarity, it's a predicate that control the evolution of segmentation process. The pixel, with the smallest difference measured this way, is allocated to the respective region. This process continues until all pixels are allocated to a region.

```
While all the pixels in image are not visited;  
  1. Choose an unlabeled pixel  $p_k$ ;  
  2. Set the region's mean to intensity of pixel  $p_k$ ;  
  3. Consider unlabeled neighboring pixels  $p_{kj}$ ;  
  If (pixel's intensity - region's mean) < threshold;  
    a. Affect the pixel  $p_{kj}$  to the region labeled by  $k$ .  
    b. Update the region's mean and go back to ③;  
  Else  $k = k + 1$  and go back to ①.  
End If  
End While
```

Figure 2. Region growing segmentation algorithm.

By assuming that the objects are localised at the center of images, the region growing segmentation is started from the corner of the image to isolate the objects in the center of the

image. The region growing segmentation algorithm used in this paper is presented in Fig.2 and an example of image segmentation is in Fig.3.

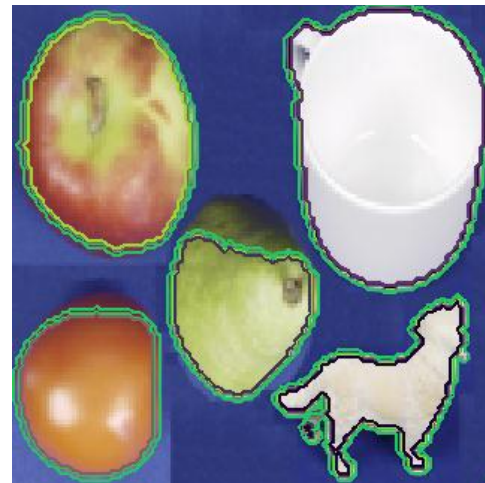


Figure 3. Example of image segmentation result.

After dividing the original image into several distinct regions that correspond to objects in a scene, the feature vector can be extracted from each region and can be considered as a representation of an object in the entire image.

The feature extraction task transforms carefully the rich content and large input data of images into a reduced representation set of features in order to decrease the processing time. Not only it enhances the retrieval and annotation accuracy, but the annotation speed as well, since a large image database can be organized according to the classification rule and, therefore, search can be performed [6].

In the feature extraction method, the representation of the image content must be considered in some situations such as: translation, rotation and change of scale. This is the reason that justifies the use of moments for feature extraction method from the segmented image.

The use of moments for image analysis and pattern recognition was inspired by Hu [7] and Alt [10]. In this paper, the moments used are:

- Hu moments.
- Legendre moments.
- Zernike moments.

A. Hu moments

For a discrete image of $M \times N$ pixels with intensity function $f(x, y)$, Hu [7] defined the following seven moments that are invariant to the change of scale, translation and rotation:

$$\phi_1 = \mu_{20} + \mu_{02} \quad (1)$$

$$\phi_2 = (\mu_{20} + \mu_{02})^2 + 4\mu_{11}^2 \quad (2)$$

$$\phi_3 = (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2 \quad (3)$$

$$\phi_4 = (\mu_{30} - \mu_{12})^2 + (\mu_{21} - \mu_{03})^2 \quad (4)$$

$$\phi_5 = (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12}) \times \left[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2 \right] + (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03}) \times \left[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right] \quad (5)$$

$$\phi_6 = (\mu_{20} - \mu_{02}) \left[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right] + 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}) \quad (6)$$

$$\phi_7 = (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12}) \times \left[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2 \right] - (\mu_{30} - 3\mu_{12})(\mu_{21} + \mu_{03}) \times \left[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2 \right] \quad (7)$$

Where,

$$\mu_{pq} = \frac{\sum_{x=0}^{N-1} \sum_{y=0}^{M-1} (x-\bar{x})^p (y-\bar{y})^q f(x, y)}{\left(\sum_{x=0}^{N-1} \sum_{y=0}^{M-1} f(x, y) \right)^{\frac{p+q+1}{2}}} \quad (8)$$

And,

$$\left\{ \begin{aligned} \bar{x} &= \frac{\sum_{x=0}^{N-1} \sum_{y=0}^{M-1} x^p f(x, y)}{\sum_{x=0}^{N-1} \sum_{y=0}^{M-1} f(x, y)} \\ \bar{y} &= \frac{\sum_{x=0}^{N-1} \sum_{y=0}^{M-1} y^q f(x, y)}{\sum_{x=0}^{N-1} \sum_{y=0}^{M-1} f(x, y)} \end{aligned} \right. \quad (9)$$

B. Legendre moments

Legendre moments were first introduced by Teague [8]. They were used in several pattern recognition applications [9]. The orthogonal property of Legendre polynomials implies no redundancy or overlap of information between the moments with different orders. This property enables the contribution of each moment to be unique and independent of the information in an image [10]. The Legendre moments for a discrete image

of $M \times N$ pixels with intensity function $f(x, y)$ is the following [13]:

$$L_{pq} = \lambda_{pq} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} P_p(x_i) P_q(y_j) f(x, y) \quad (10)$$

Where $\lambda_{pq} = \frac{(2p+1)(2q+1)}{M \times N}$, x_i and y_j denote the normalized pixel coordinates in the range of $[-1, +1]$, which are given by:

$$\begin{cases} x_i = \frac{2x - (M-1)}{M-1} \\ y_j = \frac{2y - (N-1)}{N-1} \end{cases} \quad (11)$$

$P_p(x)$ is the p^{th} -order Legendre polynomial defined by:

$$P_p(x) = \sum_{k=0}^p \left\{ \frac{(-1)^{\frac{p-k}{2}} x^k (p+k)!}{2^p k! \left(\frac{p-k}{2}\right)! \left(\frac{p+k}{2}\right)!} \right\}_{p-k=\text{even}} \quad (12)$$

And, the recurrent formula of the Legendre polynomials is:

$$\begin{cases} P_p(x) = \frac{(2p-1)x}{p} P_{p-1}(x) - \frac{(p-1)}{p} P_{p-2}(x) \\ P_1(x) = x, \quad P_0(x) = 1 \end{cases} \quad (13)$$

In this work the recurrent formula is used for calculating Legendre polynomials in order to increase the computation speed.

C. Zernike moments

Zernike moments are the mapping of an image onto a set of complex Zernike polynomials. As these Zernike polynomials are orthogonal to each other, Zernike moments can represent the properties of an image with no redundancy or overlap of information between the moments [11]. Due to these characteristics, Zernike moments have been utilized as feature sets in many applications [12].

The discrete form of the Zernike moments of an image size $M \times N$ represented by $f(x, y)$ is expressed, in the unit disk $(x^2 + y^2) = 1$, as follows [13]:

$$Z_{pq} = \frac{p+1}{\lambda} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} R_{p,q}(r_{xy}) e^{-jq\theta_{xy}} f(x, y) \quad (14)$$

Where

$$\left\{ \begin{aligned} R_{p,q}(r) &= \sum_{s=0}^{(p-|q|)/2} \frac{(-1)^s (p-s)! r^{p-2s}}{s! \left(\frac{p+|q|}{2}-s\right)! \left(\frac{p-|q|}{2}-s\right)!} \quad (15) \\ p-|q| &= (\text{even}), |q| \leq p, p \geq 0 \end{aligned} \right.$$

λ is the number of pixels located in the unit circle, and the transformed phase θ_{xy} and the distance r_{xy} at the pixel of coordinates (x, y) are [13]:

$$\left\{ \begin{aligned} \theta_{xy} &= \tan^{-1} \left(\frac{(2y-(N-1))/(N-1)}{(2x-(M-1))/(M-1)} \right) \\ r_{xy} &= \sqrt{\left(\frac{2x-(M-1)}{M-1} \right)^2 + \left(\frac{2y-(N-1)}{N-1} \right)^2} \end{aligned} \right. \quad (16)$$

Most of the time taken for the computation of Zernike moments is due to the computation of radial polynomials. Therefore, researchers have proposed faster methods that reduce the factorial terms by utilizing the recurrence's relations on the radial polynomials. In this paper, we obtained Zernike moments using the direct method.

IV. IMAGE'S CLASSIFICATION AND ANNOTATION

As the features are extracted, a suitable classifier must be chosen. Many classifiers are used and each classifier is found suitable to classify a particular kind of feature vectors depending on their characteristics.

Some of these classifiers are based on supervised learning which require an intensive learning and a training phase of the classifier parameters (parameters of Support Vector Machines [14], Boosting [15], parametric generative models [16], decision trees[17], and Neural Network[18]). They are also known as parametric classifiers. Other classifiers base their classification decision directly on the data, and require no learning and training of parameters. These methods are also known as nonparametric classifiers. The most common nonparametric classifier, based on distance estimation, is the Nearest Neighbour classifier.

There are several types of the classifier. Based on their ability to detect complex nonlinear relationships between dependent and independent variables, the multilayer neural networks are used in this paper to classify and annotate image content. Their performances are compared to those of the Nearest Neighbour classifier.

A. Multilayer Neural Network classifier

A multilayer neural network consists of an input layer including a set of input nodes, one or more hidden layers of nodes, and an output layer of nodes [19]. Fig. 4 shows an example of a three layer network, used in this paper, having input layer formed by m nodes, one hidden layer formed by 20 nodes, and output layer formed by n nodes. This neural network is trained to classify inputs according to target classes.

The training input data are loaded from the reference database while the target data should consist of a vector of all zero values except for the element i that represents the appropriate class.

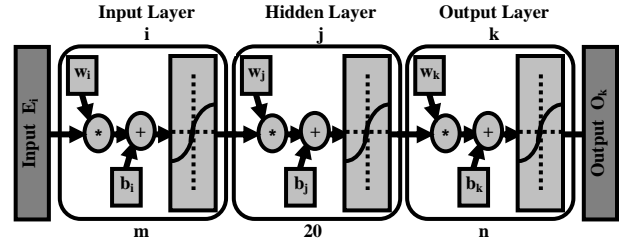


Figure 4. The three layer neural network.

After random initialisation of bias and connection weights, the training principle of the neural network is based on a loop of steps. It starts by propagating inputs from the input layer to the output layer. The error is calculated and back propagated for each layer of the neural network in order to update the bias and connection weights. When the bias and connection weights are changed, the propagation of the inputs is repeated until having the minimum error between outputs and targets. This is the criterion stop of the neural network training.

The inputs Y_i are presented to the input layer and propagated to the hidden layer using the following formula:

$$Y_j = f \left(\sum_{i=1}^m Y_i w_i + b_j \right) \quad (17)$$

Then from the hidden layer to the output layer:

$$Y_k = f \left(\sum_{j=1}^{20} Y_j w_j + b_k \right) \quad (18)$$

Finally, the outputs are:

$$O_k = f \left(\sum_{k=1}^n Y_k w_k + b_k \right) \quad (19)$$

Where f is the activation function (hyperbolic tangent sigmoid) used in the tree layer neural network, defined by:

$$f(x) = \frac{2}{1 + e^{-2x}} - 1 \quad (20)$$

At the output layer, the error between the desired output T_k and the actual output O_k is calculated by:

$$E_k = (1 - O_k^2)(T_k - O_k) \quad (21)$$

The calculated error is propagated to the hidden layer using the following formula:

$$E_j = \left(1 - Y_k^2\right) \sum_{k=1}^n w_k \cdot E_k \quad (22)$$

Then the calculated back propagation of error from the hidden layer to the input layer is:

$$E_i = \left(1 - Y_j^2\right) \sum_{j=1}^{20} w_j \cdot E_j \quad (23)$$

The bias and the connections weights of the input layer i, the hidden layer j and the output layer k are adjusted by:

$$\begin{cases} \Delta w_l = Y_l \cdot E_l \\ \Delta b_l = E_l \end{cases} \text{ for } l = i, j, k \quad (24)$$

B. Nearest Neighbor classifier

The nearest neighbour classifier is used to compare the feature vector of the input image and the feature vectors stored in the database. It is obtained by finding the distance between the prototype image and the database. The class is found by measuring the distance between a feature vector of input image and feature vectors of images in reference database. The Euclidean distance measurement is used in this paper, but other distance measurement can be also used [14].

Let X_1, X_2, \dots, X_k be the k class features vectors in the database and X_q the feature vector of the query image. The feature vector with the minimum distance is found to be the closest matching vector. It is given by:

$$d(X_q, X_j) = \min_{j \in \{1, 2, \dots, k\}} \left\{ \sqrt{\sum_i (x_q(i) - x_j(i))^2} \right\} \quad (25)$$

C. Image Annotation

For the image annotation, low-level feature vectors are calculated iteratively for each region in the image, either by using Hu moments or Legendre moments or Zernike moments. These features vector are presented either to the nearest neighbour classifier or to the Multilayer neural network that was already trained.

When features vector are presented to the nearest neighbour classifier to test matching with the feature values in reference database, the label or keyword of image class with the minimum distance is selected.

When features vectors are feed to the input layer of the Multilayer neural network, where each of the input neurons or nodes correspond to each element of the features vector, The output neurons of the neural network represent the class labels of images to be classified and annotated. Then, each region is annotated by the corresponding label that is found by the classifier.

The input layer of the neural network has a variable number of input nodes. It has seven input nodes in the case of

adoption of Hu moments, nine input nodes when adopting Zernike moments and ten input nodes when used Legendre moment as feature extraction method. However, the number of input nodes for the neural networks can be changed and increased when using Zernike and Legendre moments, as a feature extraction method, to increase the accuracy of the annotation system. The features can be also combined together and feed to the Multilayer neural network or the nearest neighbour classifier.

V. EXPERIMENTS AND RESULTS

A. Experiments

In our experiments, for each region that represent an object from the input image, the number of input features extracted using Hu invariants features extraction method is 7 (hu1, hu2, hu3, hu4, hu5, hu6, hu7) while the number of input features extracted using the order 4 of Zernike moments is 9 (Z00, Z11, Z20, Z22, Z31, Z33, Z40, Z42, Z44) and the number of input features extracted using the order 3 of Legendre moments is 10 (L00, L01, L02, L03, L10, L11, L12, L20, L21, L30). These inputs are feed to the input of the multilayer neural network or the nearest neighbour classifier in order to do matching with the feature values in the reference database. Then, the appropriate keywords are selected and used for annotation of the input image.

The fig. 5 shows some examples of image objects from the ETH-80 image database used in our experiments. The experiments are based on eight classes of objects (Apple, Car, Cow, Cup, Dog, Horse, Pears, and Tomato). The number of prototypes per class is 5.

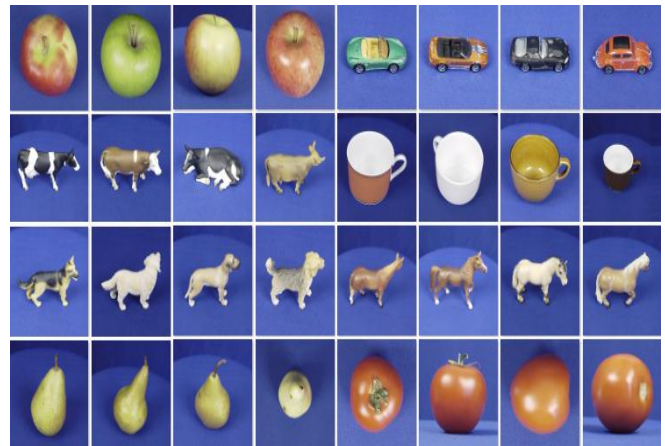


Figure 5. Some examples of objects from the ETH-80 database.

The accuracy of image annotation system is evaluated by the precision rate which is the number of the correct results divided by the number of all returned results.

All the experiments are conducted using the ETH-80 database containing a set of 8 different object images [20]. The proposed system has been implemented and tested on a Core 2 Duo personal computer using Matlab Software.

B. Results

For both of the used classifiers (the neural network classifier and Nearest Neighbour classifier), the annotation

rates of Hu, Zernike and Legendre Moments, as the feature extraction method, are given for each object in Table 1.

TABLE I. OBJECTS ANNOTATION RATE OF HU, ZERNIKE AND LEGENDRE MOMENTS USING NEURAL NETWORK AND NEAREST NEIGHBOUR CLASSIFIER

Object	Neural network classifier			Nearest neighbour classifier		
	Hu	Zernike	Legendre	Hu	Zernike	Legendre
Apple	71.20%	78.40%	92.13%	69.12%	76.84%	90.93%
Car	68.53%	74.37%	83.11%	67.03%	72.87%	81.91%
Cow	41.32%	50.00%	61.37%	39.82%	49.13%	60.10%
Cup	63.27%	71.20%	81.20%	62.07%	70.37%	80.02%
Dog	62.57%	72.30%	80.41%	60.87%	70.93%	79.10%
Horse	48.60%	56.40%	65.28%	46.80%	55.10%	63.80%
Pears	69.29%	78.40%	89.38%	66.80%	76.94%	87.85%
Tomato	67.49%	76.52%	86.40%	65.90%	75.10%	84.70%
Average	61.53%	69.70%	79.91%	59.80%	68.41%	78.55%

The general annotation rates and error rates of the neural network classifier or the Nearest Neighbour classifier based on Hu moments, Zernike moments and Legendre Moments as feature extraction method are given in Table 2. The experimental results showed that the annotation rate of the neural network classifier based on Legendre moments, Zernike moments and Hu moments is higher than the annotation rate of the Nearest Neighbour classifier based on Legendre moments, Zernike moments and Hu moments.

TABLE II. Annotation and Error Rate of Hu, Zernike and Legendre Moments using Neural Network and Nearest Neighbour classifier

Classifier	Features Extraction method	Annotation Rate	Error Rate
Neural network	Hu moments	61.53%	38.47%
	Zernike moments	69.70%	30.30%
	Legendre moments	79.91%	20.09%
Nearest neighbour	Hu moments	59.80%	40.20%
	Zernike moments	68.41%	31.59%
	Legendre moments	78.55%	21.45%

The best annotation rate is achieved when we use the Legendre moments as the features extraction method and the multilayer neural networks as a classifier. However, for the computation time, the Nearest Neighbour classifier based on Hu moments is the best.

The annotation rates of Zernike moments is lower than the annotation rate of Legendre moments because Zernike moments are computed only from pixels of an image inside a unit circle obtained by the mapping transformation over a unit circle. In one hand, those pixels outside a unit circle are not considered. On the other hand, only the absolute values of extracted complex Zernike moments are feed into the classifier.

This is the main reason for the modest results obtained by Zernike moments.

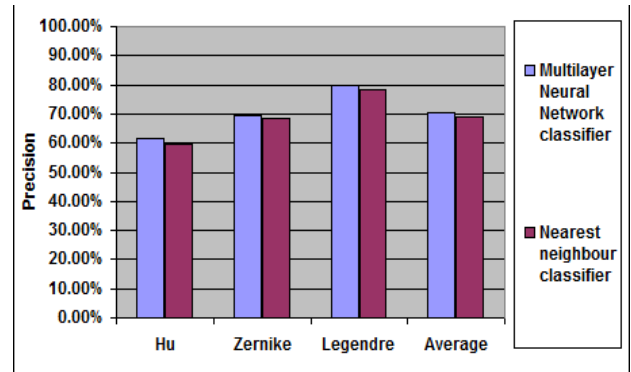
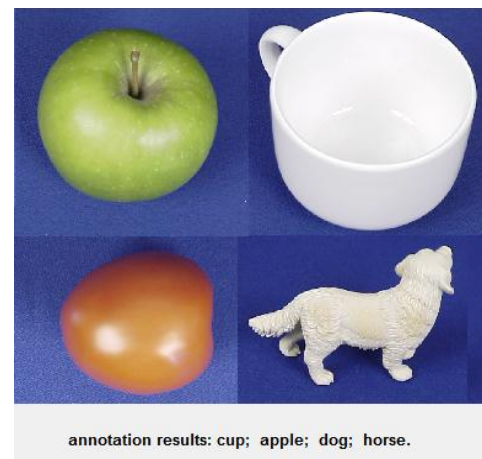
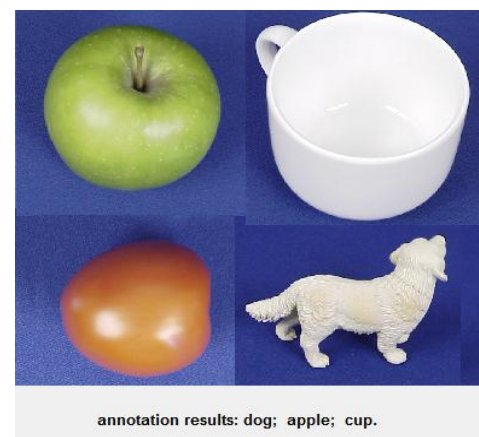


Figure 6. Comparison of general annotation rates.

The fig 6 gives the comparison of general annotation rate for each extraction method and for each classifier.



(a)



(b)

Figure 7. Example of the annotation results using Zernike moment features and (a) the multilayer neural network classifier (b) the Nearest Neighbour classifier.

The Fig. 7 presents an example of annotation results obtained using the presented system and Zernike moment as a

feature extraction method. The Graphical User Interface is illustrated in Fig. 8.

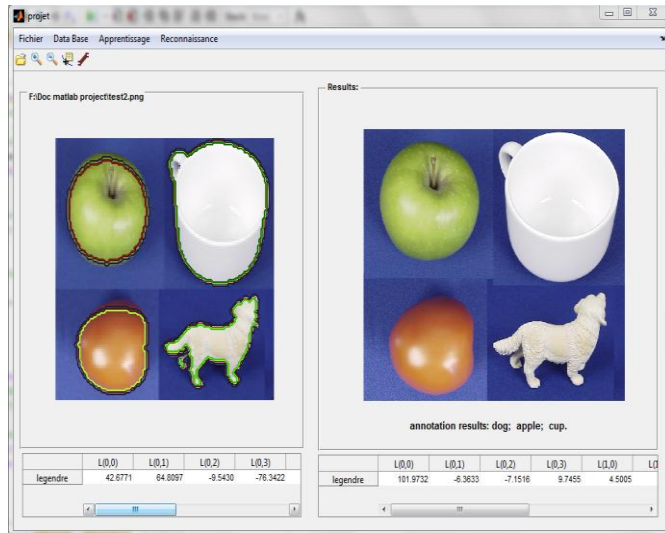


Figure 8. Graphical User Interface (GUI).

The results are also affected by the accuracy of the image segmentation method. In most cases, it is very difficult to have an automatic ideal segmentation. This problem decreases the annotation rates. Therefore, any annotation attempt must consider the image segmentation as an important step, not only for automatic image annotation system, but also for the other systems which require its use. The Multilayer neural network classifier based on Legendre moments and Zernike moments is very expensive regarding to the computation time of the features in addition to the training time of the classifier. So, any use of them in a real time for an online image annotation system will be difficult and impracticable.

VI. CONCLUSION

In this paper, we evaluated the image annotation performance of the neural network classifier and the nearest neighbour classifier based on Hu moments, Legendre moments and Zernike moments as feature extraction methods. The performance of each classifier and each feature extraction method has been experimentally analyzed. The successful experimental results proved that the proposed image annotation system based on the multilayer neural network classifier gives the best results for some images that are well and properly segmented. However, the processing time and the Image segmentation remain a challenge that needs more attention in order to increase precision and accuracy of the image annotation system. Also, the gap between the low-level features and the semantic content of an image must be reduced and considered for more accuracy of any image annotation system. Other Image segmentation and other classifiers may be considered in the future works for other image databases.

REFERENCES

- [1] N. Vasconcelos and M. Kunt, Content-Based Retrieval from Image Databases: Current Solutions and Future Directions, Proc. Int'l Conf. Image Processing, 2001.
- [2] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, Content-Based Image Retrieval: The End of the Early Years, IEEE Trans. Pattern. Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1349-1380, Dec. 2000.
- [3] Lei Ye, Philip Ogunbona and Jianqiang Wang, Image Content Annotation Based on Visual Features, Proceedings of the Eighth IEEE International Symposium on Multimedia (ISM'06), IEEE computer society, San Diego, USA, 11-13 December 2006.
- [4] Ryszard S. Chora's, Image Feature Extraction Techniques and Their Applications for CBIR and Biometrics Systems, International Journal Of Biology And Biomedical Engineering, Issue 1, Vol. 1, pp. 6-16, 2007.
- [5] Frank Y. Shih, Shouxian Cheng, Automatic seeded region growing for color image segmentation, Image and Vision Computing 23, pp. 877-886, 2005.
- [6] Zijun Yang and C.-C. Jay Kuo, Survey on Image Content Analysis, Indexing, and Retrieval Techniques and Status Report of MPEG-7, Tamkang Journal of Science and Engineering, Vol. 2, No. 3, pp. 101-118, 1999.
- [7] M. K. Hu, Visual Problem Recognition by Moment Invariant, IRE. Trans. Inform. Theory, Vol. IT-8, pp. 179-187, Feb. 1962.
- [8] M.R. Teague, Image analysis via the general theory of moments, J. Opt. Soc. Amer. 70 , pp. 920-930, 1980.
- [9] Chee-Way Chonga, P. Raveendranb and R. Mukundan, Translation and scale invariants of Legendre moments, Pattern Recognition 37, pp. 119 - 129, 2004.
- [10] F. L. Alt, Digital Pattern Recognition by Moments, J. Assoc. Computing Machinery, Vol. 9, pp. 240-258, 1962.
- [11] Sun-Kyoo Hwang, Whoi-Yul Kim, Anovel approach to the fast computation of Zernike moments, Pattern Recognition 39, pp. 2065 - 2076, 2006.
- [12] R. Sinan Tumen, M. Emre Acer and T. Metin Sezgin, Feature Extraction and Classifier Combination for Image-based Sketch Recognition, Eurographics Symposium on Sketch-Based Interfaces and Modeling, pp. 1-8, 2010.
- [13] Mustapha Oujaoura, Brahim Minaoui and Mohammed Fakir. Article: Image Annotation using Moments and Multilayer Neural Networks. IJCA Special Issue on Software Engineering, Databases and Expert Systems SEDEX(1):46-55, September 2012. Published by Foundation of Computer Science, New York, USA.
- [14] Oren Boiman, Eli Shechtman and Michal Irani, In Defense of Nearest-Neighbor Based Image Classification, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2008.
- [15] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition, In ECCV, 2004.
- [16] G. Wang, Y. Zhang, and L. Fei-Fei. Using dependent regions for object categorization in a generative framework. In CVPR, 2006.
- [17] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In ICCV, 2007.
- [18] P. Simard, D. Steinkraus, J. C. Platt, Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis, ICDAR, 2003, pp. 958-962.
- [19] S. Haykin, Neural networks: a comprehensive foundation, 2nd Edition. Prentice Hall, New Jersey. 1999.
- [20] ETH-80 database image. [Online]. Available: <http://www.d2.mpi-inf.mpg.de/Datasets/ETH80>

Minimization of Call Blocking Probability using Mobile Node velocity

Suman Kumar Sikdar

Department of Computer science & Engineering, University of Kalyani
Kalyani-741235, India

Uttam Kumar Kundu

MTech in Department of Electronics & Communication Engineering, W.B. University of Technology, India

Debabrata Sarddar (Asth Professor)

Department of Computer Science & Engineering, University of Kalyani
Kalyani-741235

Abstract—Due to rapid growth in IEEE 802.11 based Wireless Local Area Networks (WLAN), handoff has become a burning issue. A mobile Node (MN) requires handoff when it travels out of the coverage area of its current access point (AP) and tries to associate with another AP. But handoff delays provide a serious barrier for such services to be made available to mobile platforms. Throughout the last few years there has been plenty of research aimed towards reducing the handoff delay incurred in the various levels of wireless communication. In this article, we propose a received signal strength measurement based handoff technique to improve handoff probability. By calculating the speed of MN (Mobile Node) and signaling delay information we try to take the right decision of handoff initiation time. Our theoretical analysis and simulation results show that by taking the proper decision for handoff we can effectively reduce false handoff initiation probability and unnecessary traffic load causing packet loss and call blocking.

Keywords-component; Next Generation Wireless Systems (NGWS); Handoff; BS (Base Station); MN (Mobile Node); RSS (Received signal strength); IEEE802.11

I. INTRODUCTION

Handoff has become an essential criterion in mobile communication system especially in urban areas, owing to the limited coverage area of Access Points (AP). Whenever a MN move from current AP to a new AP it requires handoff. For successful implementation of seamless Voice over IP communications, the handoff latency should not exceed 50ms. But measurements indicate MAC layer handoff latencies in the range of 400ms which is completely unacceptable and thus must be reduced for wireless networking to fulfil its potential.

With the advent of real time applications, the latency and packet loss caused by mobility became an important issue in Mobile Networks. The most relevant topic of discussion is to reduce the IEEE 802.11 link-layer handoff latency. IEEE 802.11 MAC specification [1] defines two operation modes: ad hoc and infrastructure mode. In the ad hoc mode, two or more stations (STAs) recognize each other through beacons and hence establish a peer-to-peer relationship. In infrastructure mode, an AP provides network connectivity to its associated STAs to form a Basic Service Set (BSS).

Multiple APs form an Extended Service Set (ESS) that constructs the same wireless networks.

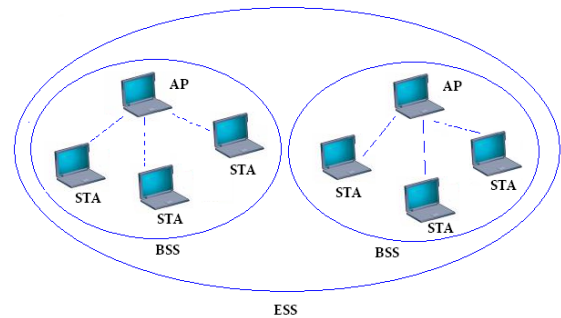


Figure 1.

A. Channel distribution

IEEE802.11b and IEEE802.11g operates in the 2.4GHz ISM band and use 11 of the maximum 14 channels available and are hence compatible due to use of same frequency channels. The channels (numbered 1 to 14) are spaced by 5MHz with a bandwidth of 22MHz, 11MHz above and below the centre of the channel. In addition there is a guard band of 1MHz at the base to accommodate out-of-band emissions below 2.4GHz. Thus a transmitter set at channel one transmits signal from 2.401GHz to 2.423GHz and so on to give the standard channel frequency distribution as shown in [Figure.2]

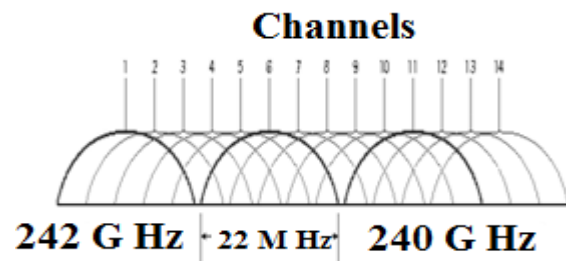


Figure 2 .

It should be noted that due to overlapping of frequencies there can be significant interference between adjacent APs. Thus, in a well configured network, most of the APs will operate on the non-overlapping channels numbered 1, 6 and 11.

B. Handoff:

When a MS moves out of reach of its current AP it must be reconnected to a new AP to continue its operation. The search for a new AP and subsequent registration under it constitute the handoff process which takes enough time (called handoff latency) to interfere with proper functioning of many applications

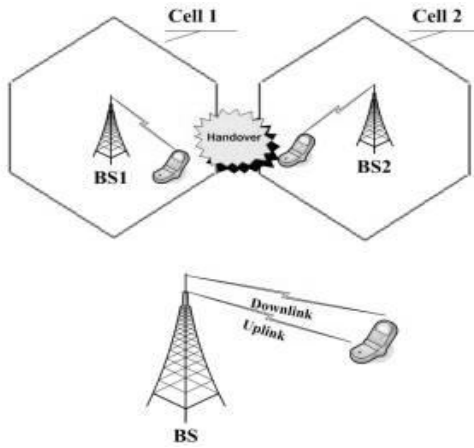


Figure2 : Handoff process

Three strategies have been proposed to detect the need for hand off[1]:

1) *mobile-controlled-handoff (MCHO):*The mobile station(MS) continuously monitors the signals of the surrounding base stations(BS)and initiates the hand off process when some handoff criteria are met.

2) *network-controlled-handoff (NCHO):*The surrounding BSs measure the signal from the MS and the network initiates the handoff process when some handoff criteria are met.

3) *mobile-assisted-handoff (MAHO):*The network asks the MS to measure the signal from the surrounding BSs. The network makes the handoff decision based on reports from the MS.

Handoff can be of many types:

Hard & soft handoff: Originally hard handoff was used where a station must break connection with the old AP before joining the new AP thus resulting in large handoff delays. However, in soft handoff the old connection is maintained until a new one is established thus significantly reducing packet loss.

The handoff procedure consists of three logical phases where all communication between the mobile station undergoing handoff and the APs concerned is controlled by the use of IEEE802.11 management frames as shown below in [fig3].

C. Scanning:

When a mobile station is moving away from its current AP, it initiates the handoff process when the received signal strength and the signal-to-noise-ratio have decreased

significantly. The STA now begins MAC layer scanning to find new APs. It can either opt for a passive scan (where it listens for beacon frames periodically sent out by APs) or chose a faster active scanning mechanism wherein it regularly sends out probe request frames and waits for responses for T_{MIN} (min Channel Time) and continues scanning until T_{MAX} (max Channel Time) if at least one response has been heard within T_{MIN} . Thus, $n * T_{MIN} \leq \text{time to scan } n \text{ channels} \leq n * T_{MAX}$. The information gathered is then processed so that the STA can decide which AP to join next. The total time required until this point constitutes 90% of the handoff delay.

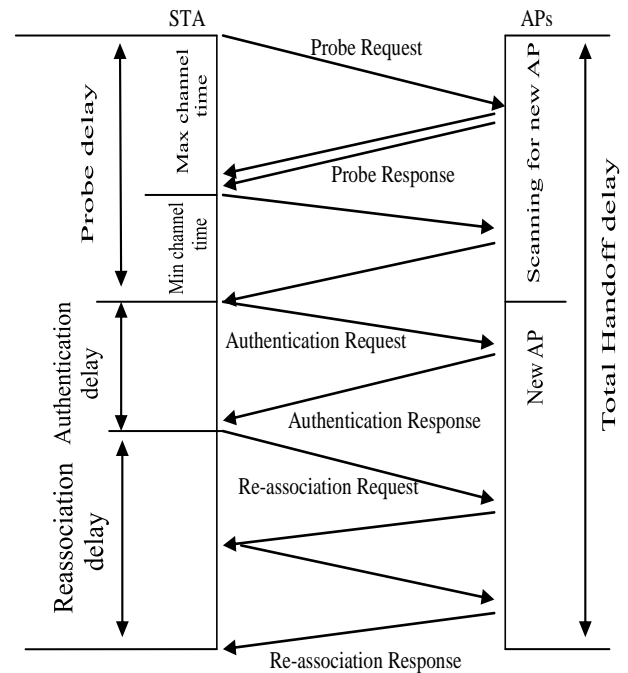


Figure 3

D. Authentication:

Authentication is necessary to associate the link with the new AP. Authentication must either immediately proceed to association or must immediately follow a channel scan cycle. In pre-authentication schemes, the MN authenticates with the new AP immediately after the scan cycle finishes. IEEE 802.11 defines two subtypes of authentication service: ‘Open System’ which is a null authentication algorithm and ‘Shared Key’ which is a four-way authentication mechanism. If Inter Access Point Protocol (IAPP) is used, only null authentication frames need to be exchanged in the re-authentication phase. Exchanging null authentication frames takes about 1-2 ms.

E. Re-Association:

Re-association is a process for transferring associations from old AP to new one. Once the STA has been authenticated with the new AP, re-association can be started. Previous works have shown re-association delay to be around 1-2 ms. The range of scanning delay is given by:-

$N \times T_{min} + T_{scan} + N \times T_{max}$

Where N is the total number of channels according to the spectrum released by a country, T_{min} is Min Channel Time, T_{scan} is the total measured scanning delay, and T_{max} is Max Channel Time. Here we focus on reducing the scanning delay by minimizing the total number of scans performed.

We have organized the rest part of this paper as follows. Sec 2 describes the related works, Sec 3 describes our proposed approach. The simulation results are given in Sec 4 and finally concluding remarks are presented in Sec 5.

II. RELATED WORKS

A lot of researches have been dedicated to enhance the performance of handover in Next Generation heterogeneous Wireless Networks. Recently a number of cross layer protocols and algorithms have been proposed to support seamless handoff in NGWS.

A number of different schemes have been proposed to reduce handoff latency in IEEE 802.11 wireless LANs. IEEE 802.11b based wireless and mobile networks [5], also called Wi-Fi commercially, are experiencing a very fast growth upsurge and are being widely deployed for providing variety of services as it is cheap, and allows anytime, anywhere access to network data.

The new age applications require a seamless handover while the small coverage of individual APs has increased the number of handoffs taking place. Thus reducing the handoff latency has become a burning issue and much work has been done to achieve this. See [6] for an overall review of popular methods suggested.

Shin et al in [7] have introduced a selective scanning algorithm with the help of channel masking technique coupled with a caching mechanism to significantly reduce the handoff delay. However, it still scans excess APs even after the new AP may have already been found and thus leaves room for further improvements.

Handoff, an inherent problem with wireless networks, particularly real time applications, has not been well addressed in IEEE 802.11, which takes a hard handoff approach [8].

In [9] the authors have introduced a novel caching process using neighbor graphs by pre-scanning neighbor APs to collect their respective channel information. The concept of neighbor graphs can be utilized in different ways and have become very popular in this field. In [10] a pre-authentication mechanism is introduced to facilitate seamless handover. [11] is a novel approach towards reducing handover latency in AP dense networks.

In [12] a cross layer handoff management protocol scheme has been proposed. They tried to enhance the handoff performance by analyzing the speed of the mobile node, handoff signaling delay, relative signal strength of old and new base station and their relation with handoff failure probability.

A novel mobility management system is proposed in [13] for vertical handoff between WWAN and WLAN. The system integrates a connection manager (CM) that intelligently detects

the changes in wireless network and a virtual connectivity manager (VCM) maintains connectivity using end-to-end principle.

Authors of [14] propose solutions towards enabling and supporting all types of mobility in heterogeneous networks. The proposed approach does not support real time applications by the network mobility functionality. This keeps the application unaware of network mobility and works as a backup for real time applications.

Handoff using received signal strength (RSS) of BS has been proposed also to reduce handoff latency in NGWS.

In [15], the authors proposed a handoff algorithm in which the received pilot signal strength is typically averaged to diminish the undesirable effect of the fast fading component. Unfortunately, the averaging process can substantially alter the characteristics of path loss and shadowing components, causing increased handoff delay.

In [16], a handoff algorithm using multi-level thresholds is proposed. The performance results obtained, shows that an 8-level threshold algorithm operates better than a single threshold algorithm in terms of forced termination and call blocking probabilities.

In [17] signal to interference ratio between old base-station and neighboring base-stations are calculated to make the handoff initiation decision for next generation wireless system or 4G networks.

In [18], a handoff algorithm using multi-level thresholds is proposed. The performance results obtained, shows that an 8-level threshold algorithm operates better than a single threshold algorithm in terms of forced termination and call blocking probabilities.

In [23], a handoff algorithm using distance measurement method is used where the distance of MN from each BS is calculated and by comparing them the best neighbor AP can be found out.

In [24], a handoff algorithm using cell sectoring method is used where the hexagonal cell is divided in 3 sectors and each sector is assigned with two neighbor APs.

In [25], a handoff algorithm using pre-scanning technique is used where the neighbor graphs are used to decide the handoff. In [26], carrier to interference ratio is calculated to find the best neighbor AP to perform handoff.

A new handoff scheme is proposed in [27] where a curve fitting equation is used to predict the direction of motion of MN and thus perform the handoff by reducing the scanning time in a considerable amount.

In [28] and [29] two schemes are proposed to reduce handoff failure probability by introducing new cell coverage area in handoff region and by using different channel allocation technique.

In [30] a vector analysis method is used to reduce the handoff latency where the direction of velocity of MN is considered as a vector quantity.

Another new handoff scheme is proposed in [31] where a tangent analysis method is used to predict the direction of motion of MN and thus perform the handoff by reducing the scanning time in a considerable amount.

III. PROPOSED WORK

In this paper we propose a handoff management protocol to support seamless handoff in Next Generation Wireless Systems. We consider the mobile node's speed, Relative Signal Strength of the base station, handoff signaling delay information and threshold distance from cell boundary to reduce false handoff initiation probability which creates unnecessary traffic load and sometimes call blocking.

For our proposed work we consider the coverage area of the base stations (BS) as regular hexagonal cells. We take two base stations into our account to explain our proposed approach, one is OBS where the call generates and other is NBS, next destination of the MN. When the MN tends to move out the coverage area of OBS it needs handoff with NBS to continue the call. Fig. 4 describes the handoff scenario between OBS and NBS. Notations used in Fig. 4 describes the following parameters.

S_{th} = The threshold value of the RSS to initiate the handoff process. When the RSS of old BS (OBS) referred to as ORSS drops below S_{th} the MN starts the HMIP registration procedures for handoff to new BS (NBS).

S_{min} = The minimum value of RSS required for successful communication between the MN and OBS.

a = The cell size.

d = The threshold distance from cell (OBS) boundary, where the MN starts HMIP registration.

We divide our proposed work into four sections:

- 1) Speed Estimation.
- 2) Measurement of threshold distance d from the cell (OBS) boundary.
- 3) RSS and S_{th} Measurement.
- 4) Handoff Management.

B. Speed Estimation:

The speed of the MN is estimated by using the Doppler Effect. The received signal frequency and the carrier signal frequency and the speed of the MN are respectively f_r , f_c , and v_m . And v_c is speed of the light. Hence

$$f_r = \frac{(v_c - v_m \cos \theta)}{v_c} \cdot f_c \dots \dots \dots (1)$$

$$\text{or, } f_r / f_c = \left(1 - \frac{v_m}{v_c} \cdot \cos \theta \right)$$

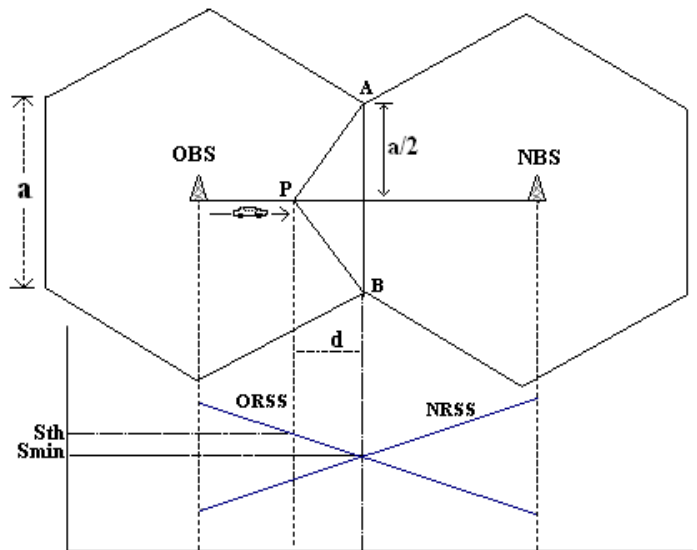


Figure 4

$$\text{or, } v_m = \frac{(f_c - f_r)}{f_c} \cdot \sec \theta$$

$$\text{or, } v_m = \left(1 - \frac{f_r}{f_c} \right) \cdot v_c \sec \theta$$

This equation helps to estimate the speed of the MN.

When the MN moves out of the coverage area of OBS through radial outward direction, $\sec \theta = 1$.

Thus

$$v_m = \left(1 - \frac{f_r}{f_c} \right) \cdot v_c \dots \dots \dots (2)$$

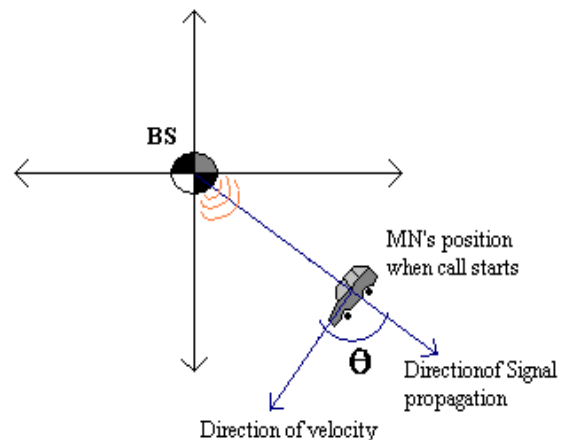


Figure 5

C. Measurement of threshold distance d from the cell (OBS) boundary:

Case1: Let handoff signaling delay is Γ , when the MN moves out of the coverage area of OBS through radically outward (Fig.5) direction the MN needs to initiate HMIP registration at the distance d from the cell boundary for a successful handoff. Otherwise the MN does not get enough time to complete the total handoff procedure.

So for a successful handoff (handoff success probability maximum) in this case

$$d \leq \Gamma \cdot v_m \dots\dots\dots (3)$$

Case2: When the MN reaches the point P, the RSS starts to drop below the S_{th} value. Here the MN needs to initiate HMIP registration. Now the MN can move to any direction from P with equal probability. For a worst case if MN moves out through PA or PB (Fig. 4), it remains within the S_{th} region and overlap region of OBS and NBS. So the MN can initiate handoff without any hesitation from here. In this case

$$PA=PB=\sqrt{(a/2)^2 + d^2}$$

For handoff

$$\Gamma < \frac{\sqrt{(a/2)^2 + d^2}}{v_m}$$

$$or, d > \sqrt{(\Gamma \cdot v_m)^2 - a^2/4} \dots\dots\dots (4)$$

In this case handoff failure probability is maximum.

Combining (3) and (4) we get,

$$\sqrt{(\Gamma \cdot v_m)^2 - a^2/4} < d \leq \Gamma \cdot v_m \dots\dots\dots (5)$$

For this condition handoff success probability lies between zero to one.

D. RSS and S_{th} Measurement:

The OBS antenna is transmitting a signal which gets weaker as The MN moves away from it. The transmitted signal power is maximum at the centre of the cell and gradually decreases as the distance from the centre increases. The received signal strength (RSS) at a distance x is given by

$$P_r = \frac{G_r \cdot G_t \cdot P_t}{(\frac{4\pi x}{\lambda})^2} \dots\dots\dots (6)$$

λ =wave length of the transmitting signal.

G_r = Receiving antenna gain.

G_t = Transmitting antenna gain.

P_t =Transmitting power.

The threshold value of the received signal strength,

$$S_{th} = \frac{G_r \cdot G_t \cdot P_t}{(\frac{4\pi y}{\lambda})^2} \dots\dots\dots (7)$$

$$Where y = \left(\frac{\sqrt{3}}{2} a - d\right)$$

$$And d = \Gamma \cdot v_m$$

We take this value as we calculate the limiting value for d previously. After this distance the MN can starts handoff initiation. After call setup the MN periodically checks the RSS, when RSS drops below the calculated threshold value the MN tries to initiate HMIP registration.

The wave propagation in multipath channel depends on the actual environment, including factors such as the antenna height, profiles of buildings, roads and geo-morphological conditions (hills, terrain) etc. These factors cause propagation loss in the channel. Hence the received signal strength may be

$$P_r = \frac{G_r \cdot G_t \cdot P_t}{L}$$

Propagation loss L is characterized by three aspects: path loss (L_p), slow fading (L_s) and fast fading (L_f).

$$L = L_p \cdot L_s \cdot L_f$$

In a typical urban area propagation loss is measured as [11]

$$L_{pu}(dB) = 69.55 + 26.16 \log_{10} f_c(MHz) - 13.82 \log_{10} h_b(m) - \alpha[h_m(m)] + [44.9 - 6.55 \log_{10} h_b(m)] \log_{10} x(km)$$

h_b = Base antenna height

h_m =MN's antenna height

x =Distance of MN from the BS. This can be calculated easily. By using Neighbor Graph, the position of MS (x_m, y_m) and the position of BS (x_b, y_b) can be found out and from these two coordinate we can calculate the distance of MS from BS. So, the distance is:

$$X = \sqrt{(x_b - x_m)^2 + (y_b - y_m)^2}$$

$\alpha[h_m(m)]$ =correlation factor for the MN's antenna height.

Where,

$$\alpha[h_m(m)] = [1.1 \log_{10} f_c - 0.7] h_m(m) - [1.56 \log_{10} f_c - 0.8]$$

E. Handoff Management:

Handoff Initiation: When the MN is going to move out its old BS its first challenge is to estimate the right time to initiate the HMIP registration. The handoff trigger unit (Fig 6) estimates the speed of the MN and its direction of motion and signaling delay information to determine the threshold RSS (S_{th}). When the RSS drops below S_{th} the handoff trigger unit sends a handoff request to handoff execution unit to start HMIP handoff procedures.

Handoff Execution:After receiving the handoff request from handoff trigger unit the MN starts HMIP registration. Once it is completed the MN switched to the new BS. The MN keeps its HMIP registration with its old BS for a specified time. This is implemented by simultaneous binding option of HMIP protocol. The MN binds care of address (CoA) of the old foreign agent (OFA) and new FA (NFA) at the gateway

FA in case of intrasystem handoff and at the HA in intersystem handoff. Therefore, the GFA and HA forward packets destined for the MN to both the CoAs during this time interval. It may be noted that for an intersystem handoff these two CoAs may belong to two different network interfaces when the MN moves between networks of different wireless technologies. Therefore multiple interfaces of the MN can be used to reduce the ping-pong effect during an intersystem handoff.

IV. SIMULATION RESULT

In this section we evaluate the performance of our approach. For our simulation we consider a macro-cellular systems with a cell size of $a=1\text{km}$ and the maximum speed of the MN is 50 km/hour. We consider the signalling delay is 40 second and wavelength (λ) is .1 m.

Here we consider an arbitrary movement of MS as shown in [figure7]. The green line in the figure shows the trajectory path of the MS. Now we will follow our proposed algorithm.

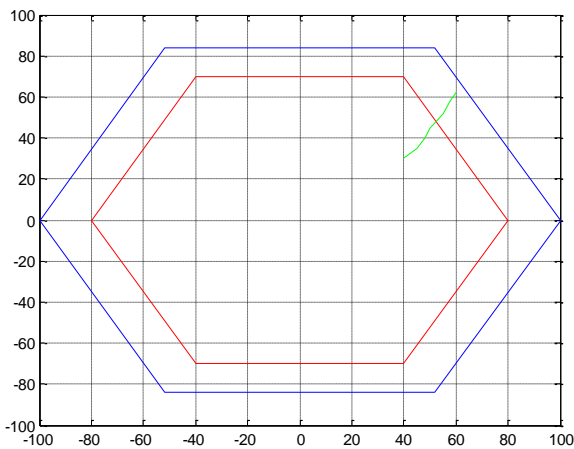


Figure 7 (*10)

The relation between distance (from the center of the cell) covered by the MN and time is shown in figure 8. Using equation (5), we can find out the range of threshold distance (d), as shown in figure 8. Again, Fig 9 shows the relation between MN's speed (km/h), MN's direction of motion (in Degree) from the place where the call was originated and the received signal frequency. The relation is established from the equation (1) using Doppler frequency shift method. The simulation result in 3D plot shows that by measuring the received signal frequency and the MN's direction of motion at different instant we can determine the MN's speed with which it is approaching towards the cell boundary.

We analyze the relationship between distance (in meter) from the Base Station and the Received Signal Strength (RSS) (in dBm) in Fig.9 using the equation (6). In figure 10, the Received Signal Strength (RSS) (in dBm) vs time (second) is shown. As the time increases, the distance of MN from the BS is increases and the RSS decreases exponentially.

After 7 seconds ($t=7$), the Received Signal Strength (RSS) goes below threshold value and handoff initiated. We simulate

the relation for three different values of carrier frequency using the parameters in Table1. This is shown in figure 11.

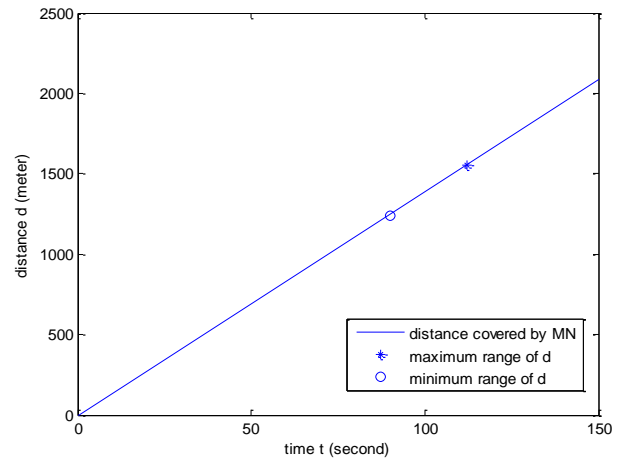


Figure 8

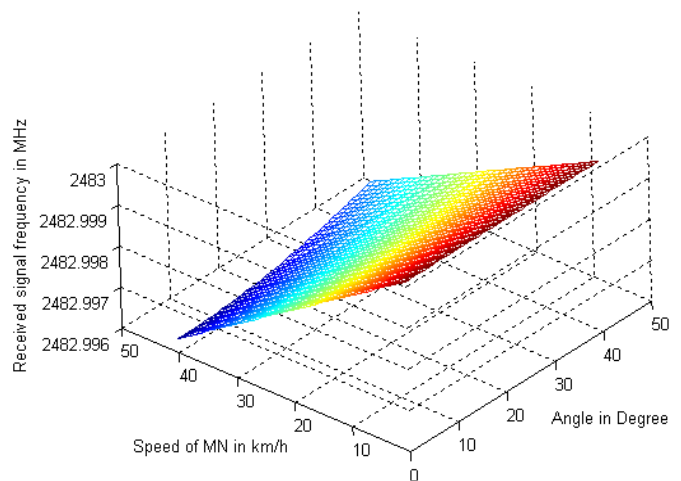


Figure 9

Table 1

Base Station antenna specification(Fiberglass Rodome, Base Station Antenna)
Frequency Range : 2.4-2.483 GHz
Gain: 18 dBi.
Max input Power/Transmitter power:115dBm/ 35 dBm
Antenna height: 30m
Polarization: vertical
Cell phone antenna specification (2.4GHz Omni Directional wireless antenna)
Frequency Range : 2.4-2.483 GHz
Gain: 7 dBi.
Max Power: 43 dBm
Antenna height: 1.6 m
Polarization: vertical

Figure 10

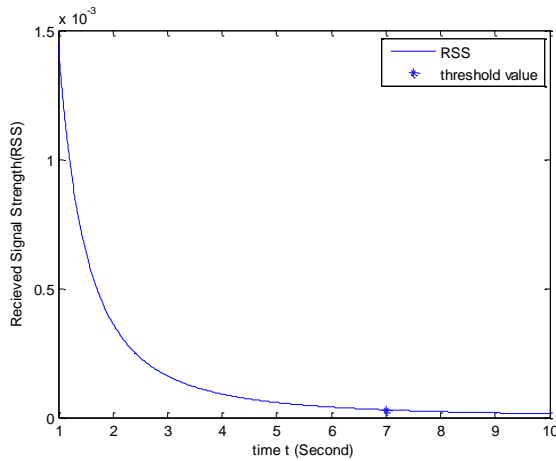


Figure 11

Fig 12 shows the relation between path-loss (in dB) and the distance from the BS (in meter). Path-loss increases exponentially as the distance increases. Results in Fig 4 and Fig 12 are correlated, they describe that when distance from the BS increases path-loss increases exponentially and as a result RSS decreases.

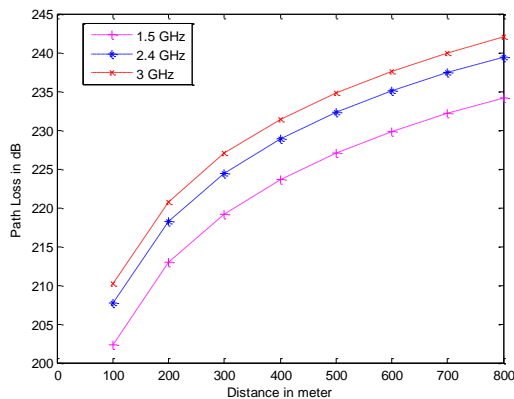


Figure 12

In figure 13, we analyze the relationship between time (in seconds) and Received Signal Strength (RSS) for three different values of MN velocity (30km/hr, 50 km/hr, 90 km/hr). MN with velocity 90 km/hr crosses the threshold value at about 1.8 second, whereas MN with velocity 50 km/hr crosses the threshold value at 7 second and MN with velocity 30 km/hr crosses the threshold value after our observation time (10 seconds). It indicates that MN with higher velocity will require handoff earlier because it loses its received signal strength more rapidly.

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.

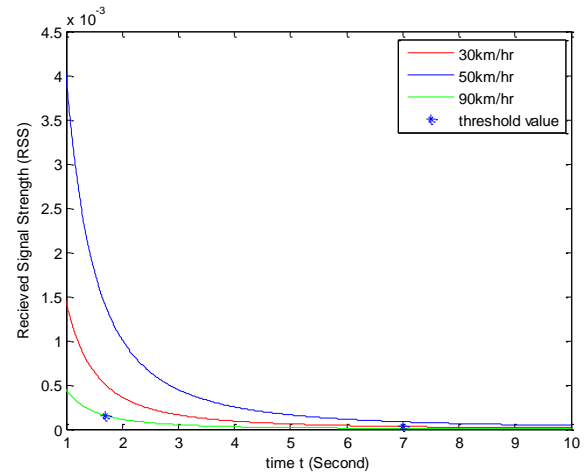


Figure 13

The plot of network traffic versus the call blocking probability is shown in the Figure 14. 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 Figure 15.

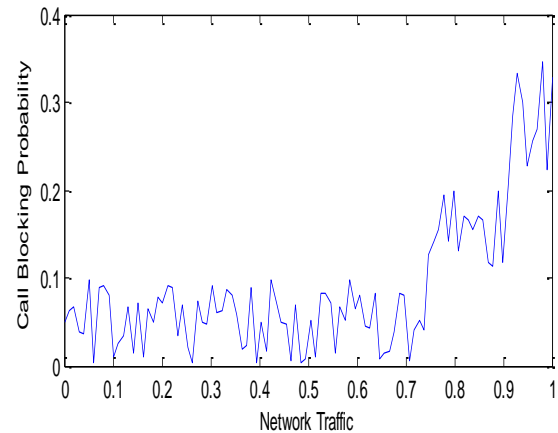


Figure 14. Plot of network traffic versus the call blocking probability

From Figure 14 it can be seen that the call blocking probability is below 0.1 for cases up to where sixty percent of the network is congested with a gradual increase, which is obvious for high congestion in the network. Figure 15 shows the significant improvement of handoff failure management against the conventional one by applying the new algorithm. Though handoff call blocking probability increases in an exponential manner with network traffic but we are able to restrict it to 8% of the maximum value which is a significant improve over the previous results. Thus our method effectively reduces the traffic blocking probability and also handoff call blocking probability.

V. ACKNOWLEDGMENT

In our article we discuss the different types of handoff procedures and associated problems for Next Generation Wireless Systems (NGWS). Then we describe our proposed

handoff management using mobile node's (MN) speed, signaling delay information and RSS.

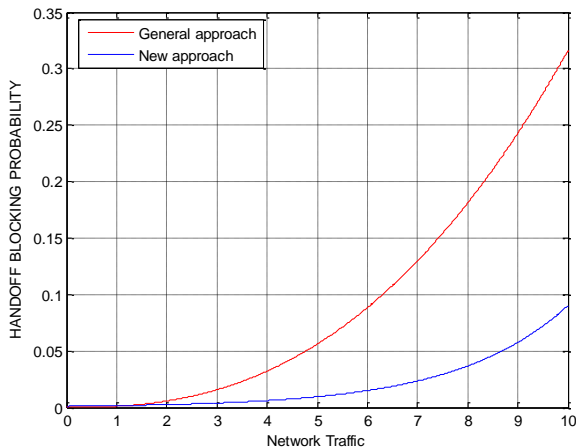


Figure 15. Plot of network traffic versus the handoff call blocking probability

We calculate the threshold value of RSS (S_{th}) and measure the distance d from the cell boundary where the MN needs to initiate the handoff procedure. Through our analysis, we observe that when the MN's speed is large and traffic load is high, it needs to initiate handoff earlier. By taking the right decision of handoff initiation time we can reduce the handoff failure probability and unnecessary traffic load resulting call termination. Our simulation results show that this approach supports the seamless handoff in Next Generation Wireless Systems (NGWS) effectively.

REFERENCES

- [1] Hye-Soo Kim, Sang Hee Park, Chun-Su Park, Jae Won Kim and Sung-Jea Ko. "Selective Channel Scanning for Fast Handoff in Wireless LAN using Neighbor Graph", July 2004.
- [2] Hongqiang Zhai, Xiang Chen, and Yuguang Fang. "How well can the IEEE 802.11 wireless lan support quality of service?" IEEE Transactions on Wireless Communications, 4(6):3084-3094, December 2005.
- [3] Yi-Bing Lin Imrich Chalmatc, "Wireless and Mobile Network Architectures," pp. 17.
- [4] AKYILDIZ, I. F., XIE, J., and MOHANTY, S., "A survey on mobility management in next generation all-IP based wireless systems," IEEE Wireless Communications, vol. 11, no. 4, pp. 16-28, 2004.
- [5] STEMM, M. and KATZ, R. H., "Vertical handoffs in wireless overlay networks," ACM/Springer Journal of Mobile Networks and Applications(MONET), vol. 3, no. 4, pp. 335-350, 1998.
- [6] Jaeyoung Choi Student Member, IEEE, Taekyoung Kwon & Yanghee Choi, Senior Member IEEE, Sangheon Pack Member, , IEEE, Fast Handoff Support in IEEE 802.11 Wireless Networks.
- [7] Arunesh Mishra, Minh Shin & William Arbaugh, An Empirical Analysis of the IEEE 802.11 MAC Layer Handoff Process. [Anshuman Singh Rawat & Henning Schulzrinne Reducing MAC Layer Handoff Latency in IEEE 802.11 Wireless LANs.
- [8] Sangho Shin, Andrea G. Forte, Context Caching using Neighbor Graphs for Fast Handoffs in a Wireless Network.
- [9] S. Park and Y. Choi Pre-authenticated fast handoff in a public wireless LAN based on IEEE802.1x mode IFIP TC6 Personal Wireless Communications. Singapore, October 2002.
- [10] Jin Teng, Changqing Xu, Weijia Jia, Dong Xuan, D-scan: Enabling Fast and Smooth Handoffs in AP-dense802.11 Wireless Networks.

- [11] Chien-Chao Tseng, K-H Chi, M-D Hsieh & H-H Chang, Location-based Fast Handoff for 802.11 Networks. IEEE COMMUNICATIONS LETTERS, VOL9, NO 4 April 2005.
- [12] S.Mohanty and I.F. Akyildiz, 'A Cross-Layer(Layer 2+3) Handoff Management Protocol for Next-Generation Wireless Systems', IEEE Transactions On Mobile Computing, Vol-5, No-10 OCT 2006.
- [13] Q. Zhang, C. Guo, Z. Guo and W. Zhu, Wireless and Networking Group, Microsoft Research Asia, "Efficient Mobility Management for Vertical Handoff between WWAN and WLAN" IEEE, communications Magazine, 2003.
- [14] I.F. Akyildiz, J. Xie, and S. Mohanty, "A Survey on Mobility Management in Next Generation All-IP Based Wireless Systems," IEEE Wireless Comm., vol. 11, no. 4, pp. 16-28, Aug. 2004.
- [15] I.F. Akyildiz, S. Mohanty, and J. Xie, "A Ubiquitous Mobile Communication Architecture for Next Generation Heterogeneous Wireless Systems," IEEE Radio Comm. Magazine, vol. 43, no. 6, pp. S29-S36, June 2005.
- [16] Brannstorm, R. Kodikara, E.R. Ahlund, C. Zaslavsky, A Mobility Management for multiple diverse applications in heterogeneous wireless networks, Consumer Communications and Networking Conference, 2006. CCNC 2006. 3rd IEEE.
- [17] K. Ayyappan and P. Dananjayan, RSS Measurement for Vertical Hnadoff in Heterogeneous Network, Journal of Theoretical and Applied Information
- [18] Leu, A. E, Mark, B. L., "Local Averaging for Fast Handoffs in Cellular Networks" IEEE Transactions On Wireless Communications, Vol. 6, No.3, March 2007, pp. 866 - 874.
- [19] Y. Kim, K. Lee and Y. Chin, "Analysis of Multi-level Threshold Handoff Algorithm", Global Telecommunications Conference(GLOBECOM'96), vol. 2, 1996, pp. 1141-1145.
- [20] S. BEN OUBIRA, M. FRIKHA, S. TABBANE, "Handoff Management In 4G Networks", IWCMC '09, June 21-24, 2009, Leipzig, Germany. ACM 978-1-60558-569-7/09/06.
- [21] Y. Kim, K. Lee and Y. Chin, "Analysis of Multi-level Threshold Handoff Algorithm", Global Telecommunications Conference(GLOBECOM'96), vol. 2, 1996, pp. 1141-1145.
- [22] D. P. Agarwal and Q. A. Zeng, 'Introduction to Mobile and Wireless Systems', chapter 3.
- [23] Distance measurement
- [24] Cell sectoring
- [25] Prescanning
- [26] Carrier to interference ratio
- [27] Curve fitting
- [28] Cahnnel allocation scheme
- [29] Introducing new cell coverage area.
- [30] Vector analysis method
- [31] Tangent analysis method

AUTHORS PROFILE

Suman Kumar Sikdar is currently pursuing his PhD at Kalyani University . He Completed his M.Tech in CSE from jadavpur University in 2011 and B-Tech in Computer Science & Engineering from Mursidabad College of engineering and technology under West Bengal University of Technology in 2007. His research interest includes wireless communication and satellite communication. Email:sikdersuman@gmail.com

Uttam Kumar kundu Completed his M.Tech in ECE from WBUT in 2010 and B-Tech in E.I.E from JIS College of engineering and technology under West Bengal University of Technology in 2006. His research interest includes wireless communication and satellite communication. Email:kunduuttam@yahoo.co.in

Debabrata Sarddar (Asst. Professor in Kalyani University) is currently pursuing his PhD at Jadavpur University. He completed his M.Tech in Computer Science & Engineering from DAVV, Indore in 2006, and his B.Tech in Computer Science & Engineering from Regional Engineering College(NIT), Durgapur in 2001. His research interest includes wireless and mobile communication Email:dsarddar@rediffmail.com.

A Pheromone Based Model for Ant Based Clustering

Saroj Bala
AKG Engineering College,
Ghaziabad, U.P., India.

S. I. Ahson
Shobhit University,
Meerut, U.P., India.

R. P. Agarwal
Shobhit University,
Meerut, U.P., India

Abstract— Swarm intelligence is a collective effort of simple agents working locally but resulting in wonderful patterns. Labor division, decentralized control, stigmergy and self organization are the major components of swarm intelligence. Ant based clustering is inspired by brood sorting in ant colonies, an example of decentralized and self organized work. Stigmergy in ant colonies is via a chemical pheromone. This paper proposes a pheromone based ant clustering algorithm. In the proposed method, ant agents use pheromone to communicate the place to make the clusters, not the path as in real ants. The pheromone concentration helps to decide pick and drop of objects with less parameters and calculations. The simulations have shown good and dense cluster results.

Keywords-Ant colony optimization; Clustering; Stigmergy; Pheromone.

I. INTRODUCTION

Swarm intelligence is an area of artificial intelligence based on the working of nature swarms like ants, honey bees, bird flocks, fish schools etc. These swarms are made up of simple agents interacting locally but emerging global patterns. An agent is a subsystem that interacts with its environment consisting of other agents but acts relatively independently from all other agents. The agent does not follow commands from a leader [11]. Ant colony optimization is the most popular swarm. It has been applied to vast variety of optimization problem. It is inspired by collective behavior of ants that are doing complex tasks without any leader. In the early 1990s, ant colony optimization was introduced by M. Dorigo and colleagues for the solution of hard combinatorial optimization problems [1]. The inspiring source of ACO is the foraging behavior of real ants.

When searching for food, ants initially explore the area surrounding their nest in a random manner. As soon as an ant finds a food source, it evaluates the quantity and the quality of the food and carries some of it back to the nest. During the return trip, the ant deposits a chemical pheromone trail on the ground. The quantity of pheromone deposited, which may depend on the quantity and quality of the food, will guide other ants to the food source. Indirect communication between the ants via pheromone trails enables them to find shortest paths between their nest and food sources.

Ant based clustering is inspired by brood sorting in ant colonies. Brood sorting shows the decentralized and self organized behavior of ants. Our work is based on combination of this behavior and the stigmergy via the pheromone. We propose here a new ant based clustering method that uses pheromone to mark the places where clusters will be formed.

II. BASIC ANT CLUSTERING METHOD

Firstly Deneubourget. al proposed a basic ant clustering algorithm [2]. He focused on clustering objects by using a group of real-world robots. In his model, the ants would walk randomly on the workspace, picking or dropping one data element from it. The ants possessed only local perceptual capabilities. They could sense the surrounding objects were similar or not to the object, they were carrying. Based on this information, they would perform the pick or drop action. The less objects of the same sort there are in the immediate environment, the greater the probabilities that it will pick it up, as given by the following function:

$$P_{pick} = \left(\frac{K^+}{K^+ + f} \right)^2 \text{ where } f \text{ is an estimation of the}$$

fraction of nearby points occupied by objects of the same type, and K^+ is a constant. The probability thus decreases with f , from 1 (when $f=0$), to $1/4$ (when $f=K^+$), and less as f tends to 1. The more objects of the same sort there are in the immediate environment, the greater the probability that it will do so, as given by the following function:

$$P_{drop} = \left(\frac{f}{K^- + f} \right)^2$$

where f is as same and K^- is a constant. The probability thus increases with f , from 0 (when $f=0$) to $1/4$ (when $f=K^-$), and more as f tends to 1.

This method was further extended by Lumer and Faieta [3] and Gutowitz [4].

III. PHEROMONE BASED METHODS IN LITERATURE

Stigmergy, the indirect communication, between the ants via pheromone trails enables them to find shortest path between their nest and food sources. In ant based clustering, many concepts are proposed for pheromone laying. Ngenkaew proposed multiple pheromones [8] in ant based clustering. Trailing pheromone is used to lead ants to return to the nest and foraging pheromone to find the new food source.

Zhang [10] proposed a method which decides the ant movement by using pheromone. It uses pheromone intensity to decide the moving orientation instead of using memory. The ant's transition probability from one place and orientation (a_1, b_1) to the next place and orientation (a_2, b_2) depends on pheromone concentration and current orientation. Before moving to the next grid, the ant will first calculate the pheromone concentration of the eight grids in the surrounding

environment, if the pheromone concentration of the eight grids is the same, the ant will move towards the original orientation. Due to the grid space of high pheromone concentration, the random moving of ants is avoided that results in acceleration of the convergence of the algorithm. The state of an individual ant has been expressed by a phase variable containing its position r and orientation θ .

Ghosh [6] proposed a method of aggregation pheromone. The ants move with an aim to create homogenous groups. The amount of movement of an ant towards a point is governed by the intensity of aggregation pheromone deposited by all other ants at that point. This gradual movement of ants in due course of time results in formation of groups or clusters. The amount of pheromone deposited by an ant at a particular location depends on its distance from it. Less is the distance; more is the deposition of pheromone. Thus aggregation pheromone density at a particular location depends on the number of ants in its closer proximity. More is the number of ants in its closer proximity; the higher is the aggregation pheromone density. The ants are allowed to move in the search space to find out the points with higher pheromone density. The movement of an ant is governed by the amount of pheromone deposited at different points of the search space. More the deposited pheromone more is the aggregation of ants. This leads to the formation of homogenous groups or clusters. The number of clusters so formed might be more than the desired number of clusters. So as to obtain the desired number of clusters, in the second step agglomerative average linkage algorithm is applied.

Liu [7] proposed an incremental method where each agent computes the information entropy or pheromone concentration of the area surrounding it, and clusters objects by picking up, dropping and moving.

Ramos [9] modified the ant-based clustering by changing the movement paradigm. His ants would move according to a trail of pheromones left on clustering formations. This would reduce the exploration of empty areas, where the pheromone would eventually evaporate. This algorithm was called ACLUSTER.

Chircop [5] proposed Multiple Pheromone Algorithm for Cluster Analysis. Ants detect individual features of objects in space and deposit pheromone traces that guide towards these features. Each ant starts off by looking for a particular feature but they can combine with ants looking for other features if the match of their paths is above a given threshold. This enables ants to detect and deposit pheromone corresponding to feature combinations and provides the colony with more powerful cluster analysis and classification tools.

IV. PROPOSED METHOD

The existing methods for ant based clustering are mainly based on memory. Whatever an ant finds in its workspace it stores in memory and the pick/drop decisions are based on the probability calculated from the memory. Few methods have used pheromone concept from real ants to change the random walk into biased walk according to pheromone.

The proposed method takes into consideration the random walk for movement and pheromone concept to avoid the memory. According to this the ants will move randomly in the

workspace, a 2-D grid, but the pick and drop actions will depend upon the pheromone on a particular cell. Pheromone on a particular cell will depend upon the objects on that particular cell and the objects lying on its neighbor cells. Here, we consider total 9 cells, a cell and its 8-neighbor cells.

The pheromone on a particular cell will be incremented by a fixed amount for every similar object in the surrounding and decremented by the same amount for every dissimilar object. Empty cells will not contribute to the pheromone and all the empty cells will not have any pheromone. Pick/drop actions will be done according to a threshold that changes its value with the cluster construction. An ant, if encounters an object on a cell, will pick it up if the pheromone on that cell is less than the threshold. If a loaded ant encounters an object, same as its load, it drops the object in neighborhood if the pheromone on the cell is greater than or equal to the threshold. The clustering process is presented below. It consists of three steps:

A. Initial pheromone laying:

This is initialization step. Every location (i, j) with an object on the grid will be assigned a pheromone τ_{ij} based on the surrounding. Let $\Delta\tau$ be the amount of pheromone change. The presence of similar objects in the surroundings increases the pheromone trail on the location by $\Delta\tau$ and a dissimilar object decreases the trail by $\Delta\tau$.

B. Cluster construction:

Ants move randomly on the grid. If an unloaded ant meets an object and finds pheromone on that location below the threshold value, it picks it up. If loaded ant comes to a location with pheromone value greater than the threshold and its load matches with the object on that location, it drops in neighborhood of location with $P_{\text{drop}} = 2\tau_{ij}$ probability.

C. Pheromone updation:

On a pick/drop action, the pheromone on that location and the surrounding location will be updated. On Pickup, $\tau_{ij} = 0$ and pheromone in the surrounding cells containing the similar object will be decreased and containing dissimilar objects will be increased. On Drop, $\tau_{ij} = \Delta\tau$ and pheromone in the surrounding cells containing the similar object will be increased and containing dissimilar objects will be decreased.

Steps B and C will alternate until final clusters form.

The number of clusters formed will be influenced by the initial pheromone assigned to a cell which in turn is dependent on the random distribution of data on the 2-D grid. The random distribution on adjacent cells on two different locations may result in two clusters having the similar elements. To merge these clusters, we propose a concept similar to agglomerative ants [12]. Each cluster will be represented by an ant. The ants will merge with the ants carrying clusters of similar elements resulting in the final clusters.

V. EXPERIMENTAL RESULTS

The proposed method has been implemented in MATLAB. Different studies have been carried out with different number of objects and ants on 2-D grid. One experiment is done on 30×30 grid. We took 100 objects of two different types, 50 objects of 'o' type and 50 objects of '□' type. Number of ants

taken is 20. The random distribution of objects and clustering results after 3000 iterations are shown in Fig. 1. The other experiment is shown in Fig. 2.

Here, we have taken 200 objects on 40×40 grid with 40 ants. All the objects are of ‘‘o’’ type, half in blue and half in red color. In the simulations, the initial threshold is taken as 0.1. Later the threshold value has been updated by $\Delta\tau$ (= 0.01 here) with cluster construction steps.

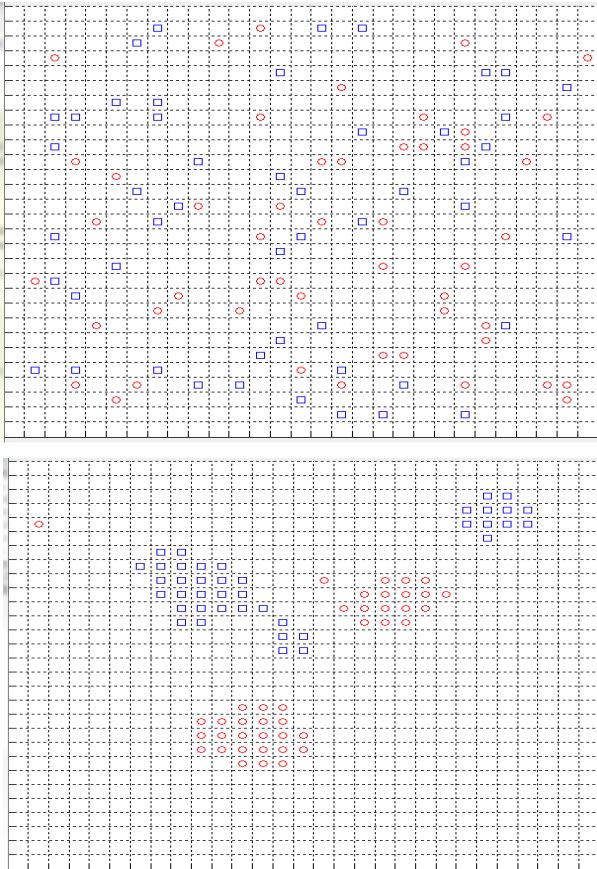


Fig. 1: Random distribution of 100 objects on 30×30 grid and clusters after 3000 iterations

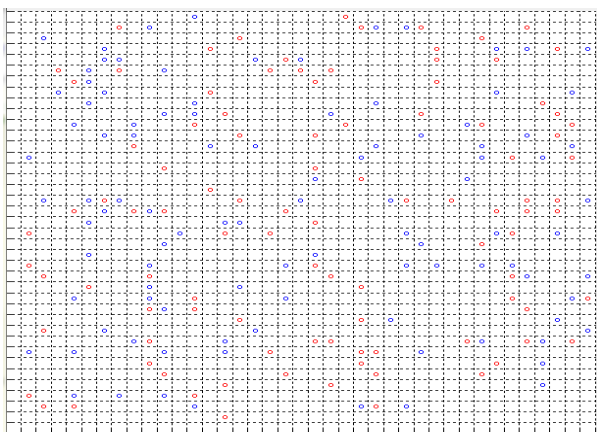


Fig. 2: Random distribution of 200 objects on 40×40 grid and clusters after 4000 iterations

VI. CONCLUSION

A new pheromone based method has been proposed for ant based clustering. Locations are carrying pheromones depending on the density of similar objects in the surrounding. The concentration of pheromone helps ant agents to pick and drop objects from a location. The simulations have shown very good and dense clusters. The ants now don't have any memory and need not to store the information of all the objects recently met during its walk. The proposed method needs very less parameters and less calculation. The pheromone will be stored only on the locations equal to number of objects and will be manipulated only on pick and drop actions while in memory based actions, ants update their memory continuously and then calculate the probability accordingly.

REFERENCES

- [1] M. Dorigo, V. Maniezzo and A. Colorni, ‘‘Ant system: optimization by a colony of cooperating agents’’, IEEE Transactions on Systems, Man, and Cybernetics-Part B, Vol. 26, No.1, pp. 1-13, 1996.
- [2] J. L. Deneubourg, S. Gross, N. R. Franks, Sendova-Franks, C. Detrain and L. Chretien, ‘‘The dynamics of collective sorting: robot-like ants and ant-like robots’’, Simulation of Adaptive Behavior: From Animals to Animats, pp. 356-363, 1991.
- [3] E. D. Lumer and B. Faieta, ‘‘Diversity and adaptation in populations of clustering ants’’, Proc. of the Third International Conference on The Simulation of Adaptive Behavior: From Animals to Animats 3, pp. 449-508. MIT Press, 1994.
- [4] H. Gutowitz, ‘‘Complexity-seeking ants’’, In Proc. of the Third European Conference on Artificial Life, 1993.
- [5] J. Chircop, J. and C. D. Buckingham, ‘‘A multiple pheromone algorithm for cluster analysis’’, International conference on swarm intelligence, ICSI 2011, Cergy, France, June 14-15.
- [6] A. Ghosh, A. Halder, M. Kothari and S. Ghosh, ‘‘Aggregation pheromone density based data clustering’’, Information Sciences 178, pp. 2816–2831. 2008.
- [7] B. Liu, J. Pan and B. Mckay, ‘‘Incremental clustering based on swarm intelligence’’, Proceedings of Simulated Evolution and Learning 6th International Conference (SEAL 2006), Springer Lecture Note in Computer Science 4247 Springer-Verlag April 2006, pp. 189-196.
- [8] W. Ngenkaew, S. Ono and S. Nakayama, ‘‘Pheromone-based concept in ant clustering’’, 3rd international conference on Intelligent System and Knowledge Engineering, ISKE 2008, IEEE, pp. 308-312.

- [9] V. Ramos, P. Pina, and F. Muge, "Self-organized data and image retrieval as a consequence of inter-dynamic synergistic relationships in artificial ant colonies", in *Frontiers in Artificial Intelligence and Applications, Soft Computing Systems - Design, Management and Applications*, IOS Press, Vol. 87, ISBN 1586032976, Santiago, Chile, pp. 500-509, 2002.
- [10] F. Zhang, Y. Ma, N. Hou and H. Liu, "An ant-based fast text clustering approach using pheromone", *Fifth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2008*, IEEE, pp. 385-389.
- [11] E. Bonabeau, M. Dorigo and G. Theraulaz, "Swarm intelligence: from natural to artificial systems", Oxford university press, 1999 .
- [12] S. Bala, S. I. Ahson and R. P. Agarwal, "Agglomerative ants for data clustering", *International journal of computer applications*, Vol. 47, No. 21, June edition, 2012.

Modern and Digitalized USB Device With Extendable Memory Capacity

J. Nandini Meeraa, S. Devi
Abirami, N. Indhuja, R. Aravind
Dept of CSE, Final year
NSIT, Salem, India

C. Chithiraikkayalvizhi
Dept of ECE, Final year
Sri Krishna College of Engineering
Coimbatore, India

K. Rathina Kumar
Assistant Prof, Dept of ECE
Knowledge Institute of Technology
Salem, India

Abstract—This paper proposes a advance technology which is completely innovative and creative. The urge of inventing this proposal lies on the bases of the idea of making a pen drive have an extendable memory capacity with a modern and digitalized look. This device can operate without the use of a computer system or a mobile. The computerized pen drive has a display unit to display the contents of the pen drive and in-built USB slots to perform data transmission to other pen drives directly without the use of the computer system. The implementation of the extendable memory slots to the computerized pen drive makes it a modern digitalized and extendable USB device. The implementation of an operating system and a processor to the pen drive are the main challenges in this proposed system. The design of the system is done in such a way that the device is cost efficient and user friendly. The design process and the hardware, software structures of this modern and digitalized USB device with extendable memory capacity are explained in detail in this paper.

Keywords-extendable memory capacity;in-built USB slots; computerized pen drive; operating system;processor.

I. INTRODUCTION

This paper primarily focuses on the innovation of a new gadget which is completely useful for storage and data transfer. The core subject of this proposal is a pen drive device. The major reason for choosing it is in-spite of developments of many advanced technologies in the field data exchange and transmission mainly like internet through which we can send and receive a lot of data and files, but the usage of the pan drive devices is still dominant through the world and very common in the student and working community. It is rare to see a person making use of the computer system even in personal or official working environments without owning a pen drive in the present scenario.

There are also so many latest devices or the mediums in which storage of data can be made, for example the cloud computing offers a way in which we can store the data in the cloud area being provide for us. The Imagination of a normal pen drive with the display gives an attractive and enables to know the details of the contents inside the pen drive. The implementation of the in-built USB ports in the body of the pen drive provides a medium, through which another pen drive being connected to it directly and data transmission can be done even without the use of the computer system.

The way in which the pen drive models are categorized is normally based on the memory size like 2GB, 4GB up to 64GB

in today's market. The main idea of this paper is change this trend and bring into the advanced concept of the adjustable memory capacity pen drive devices.

This process allows the pen drive to have extendable slots through which additional memory cards can be inserted in order to increase the size of the memory capacity of the computerized pen drive. These are the brief introduction about the modern and digitalized USB device with extendable memory capacity and the concept implementation are described as follows.

II. NORMAL PEN DRIVE MODELS AND ITS FEATURES

A. Look and Dependent on computer systems to operate

The look of the normal pen drives which are being present now has no special features are as shown as follows in the figure 1. There is no display unit or in-built USB port for data transmission. It cannot operate without a computer system; completely depend on the computers for working with it.



Figure 1. Look of a Normal Pen Drive.

B. Fixed Memory capacity

The normal pen drives are of fixed memory size, all the models are limited to a constant memory space. The memory capacity is a major issue while in need of extra space for storage and other operations being done through it.

So this paper provides a complete solution for this limitation and has a new approach over the sales of the pen drive which will not be on the memory size as being now. The variation in the memory can be done when there is a need for it.

III. COMPUTERIZED PEN DRIVE

The look of the normal pen drives which are being present now has no special features are as shown as follows in the figure 1. There is no display unit or in-built USB port for data transmission. It cannot operate without a computer system; completely depend on the computers for working with it. Computerized Pendrive is a data storage device like text, image, video etc. and it was invented by invented by Amir Ban, Dov Moran and OronOgdan and it was established by IBM in 2000.

It is known as Universal SerialBus (USB) interface and it does what a floppy disk dose. The essential components are A USB plug, microcontroller, flash memory chip, Crystal oscillator, Jumpers, LEDs and Write-protect switches. It supports the operating system like UNIX, LINUX, MACOS and WINDOWS.

Its main features are touch screen, two USB slots and has charge terminal it does what the entire computer does. We can transfer the data from one pen drive to the computerized pendrive and edit the data's to our needs. It is smaller, faster, cheaper and portable. It has the charge terminal to charge as like laptop's etc. They have a flash memory that is lower conception of power by advanced microprocessor .We can even play songs and videos.

The transformation of large data is done in less time. Then the appearance is very attractive and colorful. They are having a file system like,

- Defragment.
- Even Distribution.
- Hard Drive.

It is used to store the data's of booting a system ,computer forensics and law enforcement, booting operating system, windows vista and windows 7ready boost, audio player, media marketing and storage, arcades, brands and product promotions, operating system installation, medias, graphics, security systems and backup's. It is the portable devices like tape, floppy disk, optimal media, flash memory cards, external hard disk, obsolete devices, encryption and security. This computerized pendrive have a security code that we can keep a number security lock that no other people can access it. It has a touch screen ability to perform its job and know the contents of the memory in the pen drive.

There is a need of behavior analysis of flash-memory storage systems and other USB devices for storages and their evaluations. In particular, a set of evaluation metrics and their corresponding access patterns are proposed. The behaviors of flash memory are also analyzed in terms of performance and reliability issues [1].

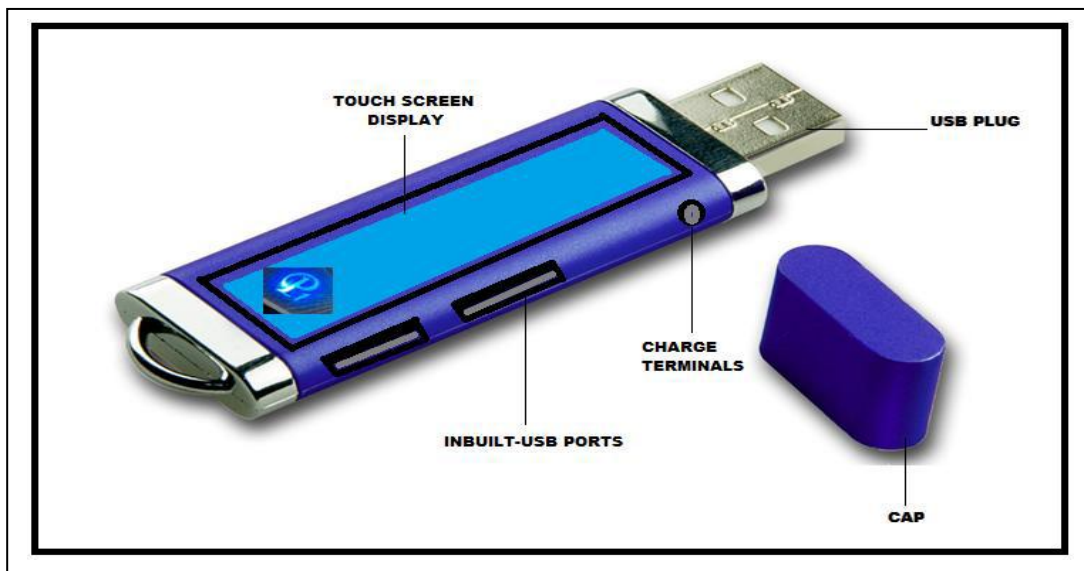


Figure 2.The computerized pen drive.

The security of the pen drive can be implemented be some techniques which can identify specific drive. We can describe the methods of digital evidence analysis [2] about USB thumb drives or devices such as Computerized Pen Drive. The rate of the pen drives which is implemented with these features can also be cheap as the normal drives being in the market now. This can really create a dynamic storage and transmission device. This proposal when implemented has more impact on the pen drive sales. The advantages of using such USB pen drives and flash drives are mainly because their power

consumption and energy overhead features [3]. The main advantages are listed as follows:

- Very handy and compact
- Display unit to display its contents
- Extendable memory slots
- High speed processor and user friendly OS.
- Chargeable battery
- Affordable cost.

A. Display unit

The display unit is the advanced feature in the modern pen drives. The display of the pen drive can be designed using touch screen. Using this unit we can see the contents in the pen drive using the touch screen. Using the touch screen in the pen drive will reduce the space occupied by the buttons in the pen drives. There are many kinds of display technologies. The storage tube graphic display is the oldest technology used for display but now we are using the new virtual therapy for modern pen drives. The quality of the display can be increased by using the technology that we are using for the display. The quality of the display will be varying based on the technology we are using. We have to consider the various factors such as quality and speed while choosing the technologies that are using for display.

B. In-built USB Port

In the computerized pen drive there will be In-built USB port slot. Using that slot you can insert two pen drives in the space provided for it. This computerized pen drive is working as a computer system by providing all the features.

C. Data Transmission

The Data transmission is also the advanced feature in the computerized pen drive. Using this feature we can transmit the data to any other USB devices and the transmission from any other USB devices can be made very easy and simple using this computerized pen drive. The working of the computerized pen drive is similar to the computer system. The data transmission through this computerized pen drive will reduce the consumption of power. The need to connect portable electronic devices to each other has accelerated the adoption of USB on-the-go as an industry standard wired interface [4] for interfacing the two devices concepts. The data transmission in the Computerized Pen Drive can also be done without the use of the system.



Figure 3. Data Transmission

D. NIDR Processor

The use of NIDR processor is used to increase the speed and the data can be transmitted in a more speed than any other processor [5]. First it will fetch the instruction and send to the identifier unit to identify the type of instruction. The multiple instruction queues are implemented to increase the temporary memory. Based on the type of instructions all the instructions will be decoded at a time and finally those instruction are send to the execute unit for execution. The operation of this processor is shown as following.

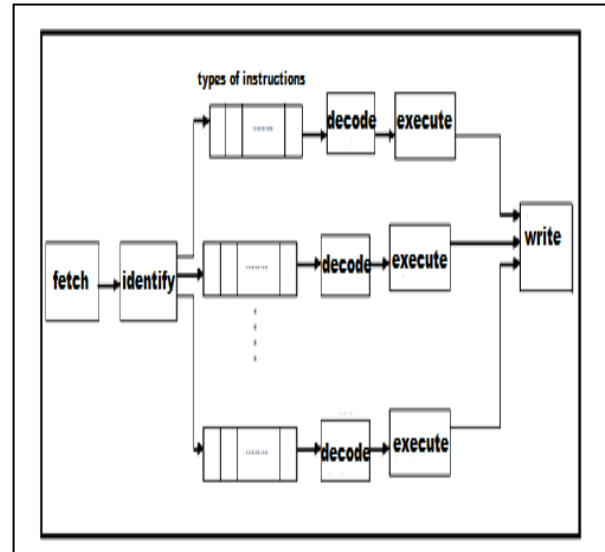


Figure 4. NIDR Processor

E. Operating System

Operating system is very important in this pen drive, because without the support of pen drive we cannot display the contents in the pen drive through display unit. The other main use of operating system in the computerized pen drive is to maintain the formats of contents in the pen drive and the transmission of contents from the computerized pen drive. The use of operating system in the pen drive will also provide the user friendly environment so that the user will feel more comfort and easy while using this computerized pen drive. We can also use the Caernarvon operating system demonstrate that a high assurance system for smart cards was technically feasible and commercially viable. The entire system has been designed to be evaluated under the Common Criteria at EAL7 [6]. The processor must also be implemented for the functioning of the Computerized Pen Drive. We use the Nurture IDR segmentation and multiple instructions queues in superscalar pipelining processor [7], which is very fast and the efficient processor.

F. Charge Terminals

While using the computerized pen drives if the battery is low we can charge the pen drive by using the mobile phone adapter in the charge terminal slot. It requires only small amount of power. In the Computerized Pen Drive the charge terminals are optional while constructing of the pen drive.

IV. DESIGNING EXTENDABLE MEMORY SLOTS

The core idea of this proposal is creating the extendable memory slots in the body of the pen drive. This can be used in case there is a need of extra memory space. The memory can be extended by adding the memory cards in the slots provided therefore increase the space for storage. By implementing this concept the sales or the division of the pen drive devices can be brought into end.

A. Reasons for Implementation of Extendable Memory

The two main reasons which made us to design the concept of extendable memory slots in the pen drive devices are being explained as follows.

1) *Insufficient memory space in the pen drive:* Suppose a person is working a pen drive of memory capacity of 2GB and want to store and transmit a digital file of size more than 2GB may be of 3GB. In this case a memory card of 2GB can be inserted into the pen drive in the memory slots available and the capacity of the pendrive can be extended to 4GB and used for the storage of the data.

2) *Enabling pen drive to act as a card reader:* If the pen drive has the in-built slots for the memory cards then when the card is inserted the contents of the memory card can be read through the pen drive itself. The pen drive can act as a card reader also.

B. Memory Card

A memory card can also be called as a flash card and it is basically a type of electronic storage device based on flash memory concept for storing digital information like films, song and photos. The Memory can be used in many electronic devices like mobile phones, digital cameras and in MP3 players. It can also be used in laptop and desktop computers. The memory cards are basically small in size and rewritable storage device. The data which are stored in it will be retained even without power.



Figure 5.Types of memory cards

There are many types in the structure and the purpose of the memory cards, the different types of the memory card available in the market are listed as follows.

- MiniCard (Miniature Card) (max 64 MB (64 MiB))
- SmartMedia Card (SSFDC) (max 128 MB) (3.3 V,5 V)
- PCMCIA ATA Type I Flash Memory Card
- Compact Flash Card (Type I), Compact Flash High-Speed

- xD-Picture Card, xD-Picture Card Type M
- Memory Stick, MagicGate Memory Stick (max 128 MB)
- Secure Digital (SD Card), Secure Digital High-Speed, Secure Digital Plus/Xtra/etc
- MU-Flash (Mu-Card) (Mu-Card Alliance of OMIA)
- C-Flash

- SxS (S-by-S) memory card, a new memory card specification developed by SanDisk and Sony. SxS complies to the Express Card industry standard.
- NexflashWinbond Serial Flash Module (SFM) cards, size range 1 mb, 2 mb and 4 mb

C. Structure of theMemory Card Slots

The hardware space required for the implementation of this idea of creating the memory slots in the body of the pen drive does not occupy much space. The size of the memory card is very less. The following figure shows the structure of the memory card slots in the body of the pen drive. The inclusion of this particular feature in the pen drive makes the computerized pen drive a complete advanced model hence giving raise a new gadget for storage and for data transmission purpose. The major factor which has to be kept under consideration is the size of the pen drive device because the main advantage of these USB pen drives is that they are basically very handy and compact by implementing this feature we must not spoil the nature of the pen drives look and smart sizes. The one more important thing is the cost for implementing this features on a pen drive device must be low as possible.



FIGURE 6. MODERN AND DIGITALIZED USB DEVICE WITH EXTENDABLE MEMORY CAPACITY

V. CONCLUSION

In this paper we have tried to throw light on the idea of creating a new computerized and digitalized gadget which is extremely smart and innovative. The primary goal of this proposal is to create a new pen drive model which can have an extendable memory and also be used as the card reader at the same time. Adding few features like the display screen, an operating system with the processor using the concept Nurture IDR segmentation and multiple instructions queues in

superscalar pipelining processor, will support the transmission and also manage the contents of the pen drive and the inbuilt USB ports. The cost of the design of this pen drive is focused to cost efficient and the steps for controlling the effects against virus attack are also implemented.

REFERENCES

- [1] Po-Chun Huang, Yuan-Hao Chang, Tei-Wei Kuot, Jen-Wei Hsieh, Miller Lin. May 2008, "The Behavior Analysis of Flash-Memory Storage Systems", Object Oriented Real-Time Distributed Computing (ISORC), 2008 11th IEEE International Symposium, Print ISBN: 978-0-7695-3132-8 J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [2] Keun-Gi Lee, Hye-Won Lee, Chang-Wook Park, Je-Wan Bang, Kwonyoung Kim, Sangjin Lee, 13-15 Dec. 2008, "USB Pass On: Secure USB Thumb Drive Forensic Toolkit", Future Generation Communication and Networking Second International Conference. Print ISBN: 978-0-7695-3431-2.K. Elissa, "Title of paper if known," unpublished.
- [3] O'Brien K, Salyers D.C, Striegel A.D, Poellabauer C, June 2008, "Power and performance characteristics of USB flash drives", World of Wireless, Mobile and Multimedia Networks International Symposium.
- [4] Remple, T.B., Qualcomm, San Diego, CA, USA, June 2003 "USB on-the-go interface for portable devices", Consumer Electronics, 2003. ICCE. 2003, IEEE International Conference. Print ISBN: 0-7803-7721-4
- [5] Seiki Takahashi, Byoung Jun Lee, Jai Hyun Koh, Satoru Saito, Bong Hyun You, Nam Deog Kim, and Sang Soo Kim, June 2009 "Embedded Liquid Crystal Capacitive Touch Screen Technology for Large Size LCD Applications", SID Symposium Digest of Technical Papers.
- [6] David C. Toll, Paul A. Karger, Elaine R. Palmer, Suzanne K. McIntosh, Sam Weber, 2008 "The Caernarvon secure embedded operating system", Newsletter ACM SIGOPS Operating System.
- [7] J.Nandini Meeraa, N.Indhuja, S.Devi Abirami and K.Rathinakumar, "Nurture IDR Segmentation and Multiple Instruction Queues in Superscalar Pipelining Processor", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 1, November 2011

Enhanced Modified Security Framework for Nigeria Cashless E-payment System

Fidelis C. Obodoeze
Dept. of Computer
Science Renaissance
University, Agbani,
Enugu, Nigeria

Francis A. Okoye
Dept. of Computer
Science & Engr.
Enugu State
University of Science
& Technology
(ESUT) Enugu,
Nigeria

Samuel C. Asogwa
Dept. of Computer
Science Michael
Okpara University of
Agriculture Umudike,
Nigeria

Frank E. Ozioko
Dept. of Computer
& Information Science
Enugu State
University of Science
& Technology
(ESUT) Enugu,
Nigeria

Calister N. Mba
Dept. of Computer
Engr. Caritas
University Amorji-
Nike, Enugu, Nigeria

Abstract—In January 2012, the Nigeria Apex Bank, Central Bank of Nigeria (CBN) rolled out guidelines for the transition of Nigeria's mainly cash-based economy and payment system to cashless and electronic payment (e-payment) system ending over 50 years of mainly cash-based operated economy and payment system. This announcement elicited mixed reactions firstly excitement due to the enormous benefits this transition will impact on Nigeria economy and at the same time elicited panic due to unpreparedness of the economy to transit successfully to electronic payment in a system hitherto filled with bobby trap of security challenges. Ten months later after the introduction of the policy, only a handful of the major stakeholders are fully compliant mainly because of the complexity and the high prohibitive cost of implementation of CBN adopted security framework, the Payment Card Industry Data Security Standard (PCI DSS). This paper surveys the security challenges facing the full implementation of the cashless epayment policy of Nigeria and at the end introduced an enhanced modified security framework for Nigeria's cashless economy that may be easier and cheaper to implement by the majority of the stakeholders after studying the loopholes in the current Nigeria epayment system models.

Keywords—Cashless economy; electronic payment security; CBN; Nigeria; PoS; ATM; Bank Server; PCI DSS.

I. INTRODUCTION

In January 2012, Central Bank of Nigeria (CBN), the Nigerian apex bank rolled out guidelines for the switching of Nigeria's payment system which was hitherto largely cash-based to cashless and electronic payment system. The pilot test of this transition started in Nigeria commercial capital city, Lagos, to test run the workability of the project. By introducing this cashless epayment system in Nigeria, CBN aimed at reducing the amount of physical cash circulating in the economy and encouraging more electronic-based transactions [1].

Nine months into the pilot 'Cashless Lagos Scheme', security concerns have emerged. With Nigeria gradually transiting from cash to an electronic based economy, by virtue of the implementation of the Central Bank of Nigeria's (CBN) cashless policy, cyber criminals and hackers in the country who hitherto attacked businesses and individuals across the Atlantic have started re-directing their energies towards exploiting possible loopholes in the electronic payment system in order to

perpetuate fraud [2]. Analysts have warned financial institutions and other stakeholders in the electronic payment industry to step up investments in security of electronic transactions, or risk been overwhelmed by the spate of sophisticated cyber-attacks that would soon arise as a result of the sheer volume of financial transactions online [3]. According to a report from Symantec's latest Internet Security Threat, Nigeria has moved six positions up the ladder, to occupy the 59th position globally, amongst countries with greatest Internet Security threat. Industry analysts have expressed concerns about Nigeria's rising cybercrime profile. Accordingly, the apex bank, CBN, had directed all players in the banking and electronic payment sector to speedily attain Payment Card Industry Data Security Standard (PCI DSS) compliance which is the security standard adopted by the CBN for Nigeria cashless epayment system. This, according to the apex bank, is to ensure that card users in the country would continue to experience enhanced payment data security. But up till now, out of the 21 banks currently operating in Nigeria, only Access Bank Plc has achieved compliance [4]. It was discovered in [4] that only Interswitch and ETransactPlc among switching services providers, have achieved compliance; ValuCard being the only third-party card processing firm to achieve full compliance. The major reason for the massive non-compliance to the CBN's recommended security standard, the PCI DSS, is not far-fetched. It has been discovered that the major reasons for massive non-compliance by the major stakeholders to the PCI DSS are firstly because of PCI DSS high cost of implementation and maintenance, secondly because of its implementation complexity.

For instance, it will cost a business stakeholder (a merchant) in Nigeria about \$20,000 to fully implement PCI DSS and additional \$1000 per year for payment of software update on Electronic Point of Sale (EPOS) [5]. Also the merchant has to bear additional cost of periodic system vulnerability and compliance scan by paying a third-party firm appointed by PCI DSS operators to ensure full compliance to PCI DSS. This cost is highly prohibitive in Nigeria; in fact only commercial banks and perhaps the electronic payment switching companies can afford this cost. This explains why only Access Bank Plc and payment switching companies Interswitch, ETransact and card provider ValuCard have become fully compliant. An average Nigeria merchant does not have up to \$20,000 as its total operating capital (TOC). Apart

from the cost of implementation and maintenance, it equally costs additional expenses for a merchant or any other stakeholder to send his/her staff for training on how to implement PCI DSS.

Another major reason for non-compliance to PCI DSS in Nigeria is the issue of complexity of implementation. The PCI DSS was not designed with simplicity and user-friendliness in mind. The full implementation of PCI DSS is very complex and complicated and at such very difficult to learn and grasp by IT security professionals. Because of this reason, it costs epayment stakeholders a lot of fund and time to train their IT Desk and security officers to learn the PCI DSS implementation.

Apart from these reasons, there is equally much suspicion about the real intentions of the PCI DSS originators- the credit card companies (the VisaCard, eBay, PayPal and MasterCard). It was reported that many court cases have arisen as a result of imposition of fines/penalties on merchants by the PCI DSS originators in the USA. It was argued that fines were imposed by the originators on merchants even where there was not clear case of fraud, that the real intentions of the PCI DSS originators is to make profits from fines they impose on merchants[4].

II. CURRENT EPAYMENT ARCHITECTURE IN NIGERIA

The following stakeholders are currently involved in the Nigeria cashless and electronic payment (Epayment) system introduced by the apex bank CBN:

- Consumer or customer,
- Switching Companies (Epayment processors e.g. Interswitch and ETransact)
- Card processing companies (e.g. ValuCard)
- Merchants or retailers,
- Mobile Money providers (Mobile Payment Providers),
- Commercial Banks and
- The financial regulatory agency (the Central Bank of Nigeria)

According to [6,7], electronic payment is possible for business transactions involving:

- Business-to-business (B2B) and
- Business-to-Customer (B2C).

Business-to-business (B2B) transactions mostly involves Bank-to-Bank electronic fund transfer (ETF) and Electronic Cheque Clearing System (ECCS) while Business-to-Customer (B2C) transactions involves the use of payment terminals by consumers/customers such as Point-of-Sale(PoS) terminals, Mobile payment terminals, Mobile/Internet Banking terminal, Automated Teller Machine (ATM) terminal and Online Merchant portals/ecommerce shops.

Figure 1 depicts the illustration of Nigeria current electronic payment architecture. The electronic payment switching in all the cases as depicted in Figure 1 are handled by the switching companies (epayment processors) such as Interswitch and

ETransact while the Mobile payment providers, authorized by CBN, handle electronic payment such as ePurse, eMoney, Pocket Money etc. provisioning using the Telecommunication platforms (GSM) and electronic switching by the switching companies. The Nigerian electronic payment architecture uses about five (5) different models of payment transactions as depicted in Figures 2 to 6.

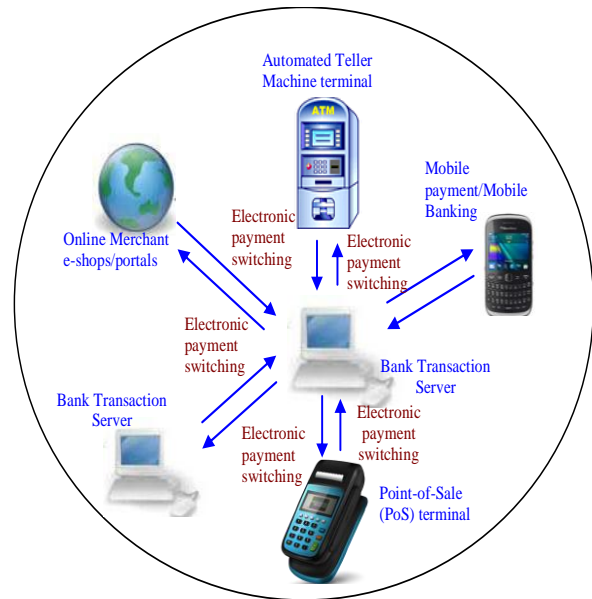


Figure 1. Nigeria's current electronic payment architecture

To solve the problems of security challenges and vulnerabilities associated with electronic transactions and payments, the epayment models of an indigenous economy must be studied and understood [8]. In Nigeria cashless electronic payment system, the epayment present electronic payment models are outlined below:

- Consumer to Merchant epayment model using PoS terminal (owned by the merchant) and bank issued credit/debit cards (owned by the consumer),
- Consumer's ATM payment model using the bank issued credit/debit cards and ATM terminal installed in banks' premises to pay customers,
- Consumer to Merchant epayment model using online portals and bank issued credit/debit cards (owned by consumers),
- Bank to Bank Electronic Payment System model using Electronic Fund Transfer (ETF) and Electronic Cheque Clearing system(ECCS), and
- Mobile Payment/Internet Banking model using consumer's mobile equipment, third-party epayment software and Bank's Transaction Server to pay consumers using telecommunication gateway and switching platform.

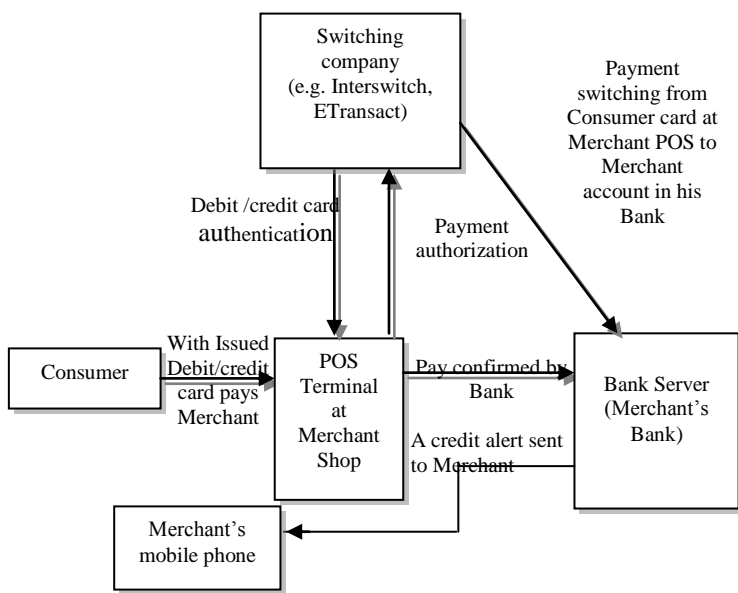


Figure 2. Epayment model using Point Of Sale (PoS) terminal to pay merchant by consumer using bank issued debit/credit cards interconnecting switching platform and Bank Server

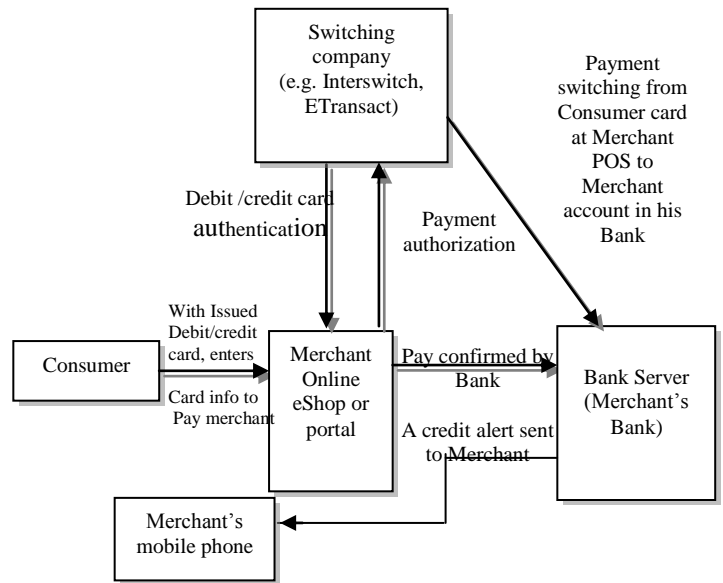


Figure 4. Epayment model using bank issued credit/debit card by consumer to pay merchant at online eCommerce shop or portal interconnecting switching platform and Bank Server

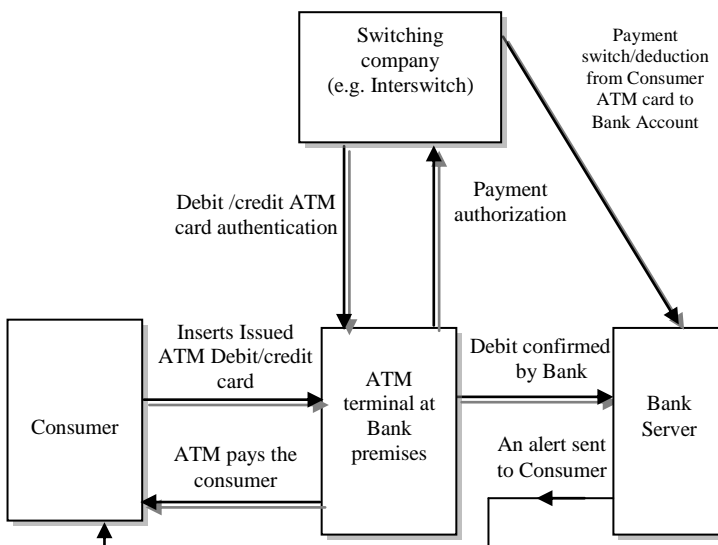


Figure 3. Epayment model using ATM terminal to pay consumer/customer having credit/debit card by Banks connecting switching company and Bank Server

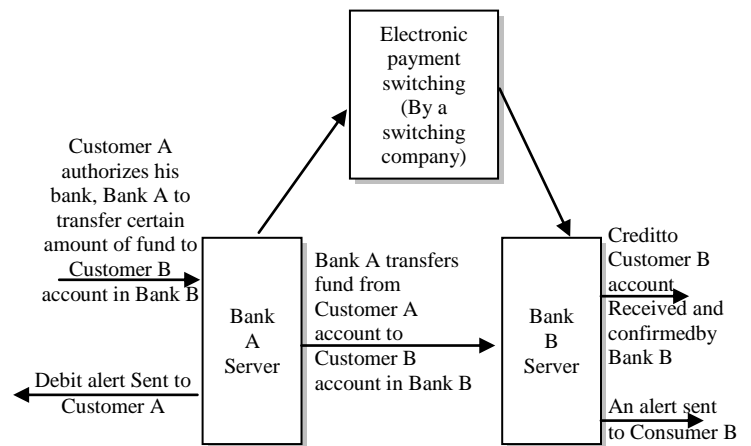


Figure 5. Epayment model involving Bank to Bank Electronic Fund Transfer (ETF) for two businesses to pay each other connecting a switching platform and two Bank Servers

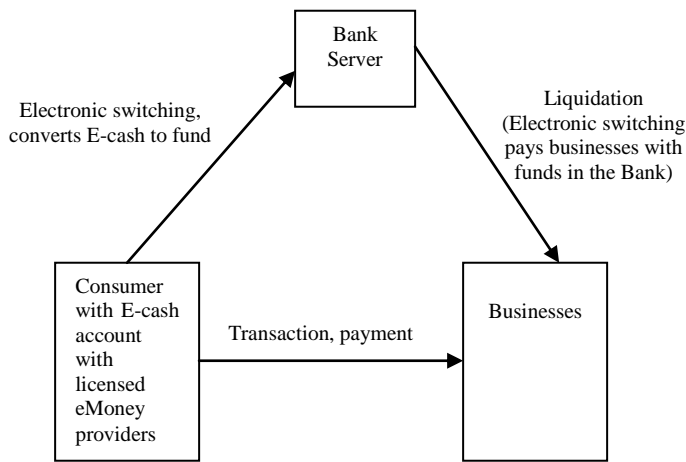


Figure 6. E-payment model using Mobile payment platform for a consumer (with eMoney account with E-Money provider) to pay a business or businesses with E-cash converted to funds in the Bank using electronic switching to switch E-cash to fund in the Bank (liquidation)

III. SECURITY CHALLENGES AFFECTING NIGERIA ELECTRONIC PAYMENT SYSTEM

The following security challenges have been identified as the key issues affecting the successful implementation of electronic payment systems in Nigeria's cashless economy:

- Identity threats,
- Fraud committed by inside-outside collusion by fraudsters and corrupt banks' employee in order to defraud the bank or their customers,
- Forgery,
- Armed robbers blowing up ATM terminals installed in Bank premises with explosives and dynamites in order to cart away physical cash,
- Cloning of credit and debit cards of bank customers' cards by hackers through various social engineering tactics to unsuspecting Nigerians to reveal their cards pin ids to enable the hackers withdraw funds from customers' bank accounts using ATM terminals.
- Online criminals and hackers exploiting security vulnerabilities in merchant websites to receive products or airline tickets for free or at reduced prices without paying a dime.
- Customers' privacy invasion and threats.
- Dearth of technical and security awareness and knowledge by most users which make them gullible to the antics of e-payment and online fraudsters.
- Lack of legal or legislative bill to sanction offenders or e-payment criminals so as to serve as deterrent.

The security challenges identified above fall into three categories of security challenges: physical, data/information and operational security challenges. Any security system designed for Nigeria's e-payment system must take into full account of these security vulnerabilities and challenges identified above as well as the loopholes identified in the

current Nigeria e-payment security models in order to be successful.

A. E-Commerce Security Elements

For a security system said to be considered adequate or successful for any electronic payment system, it must satisfy the following basic e-commerce security properties:

- Data confidentiality or privacy or secrecy,
- Data integrity,
- Non-repudiation,
- Data authentication and
- Reliability

Encryption and decryption provides the confidentiality or secrecy of data communication, but intruders still can tap transmission media and falsify the message [9]. Also reneging and forgery from the two parties of the transaction may exist. Data authentication and electronic signatures are applications of encryption systems that provide verification of message integrity, data origin authentication, and protection against repudiation. Data authentication is a procedure that allows parties to verify that the received messages are authentic, i.e., they are genuine and have come from their alleged sources. Before the message is sent, it is fed into an authentication code generator, which creates a code that accompanies the message to the receiver. Upon receipt of the message, the receiver performs an authentication code calculation and obtains another authentication code. If the message has not been altered, the same authentication code will be generated. Multiple algorithms can be used for authentication code generation. If there are no authentication keys, some kind of hashing function can be used, but this method suffers from collision problems, i.e., a falsified message generates the same authentication code with the original message. Encryption/decryption algorithms are often used for authentication. When the AES 128 bit cipher algorithm is used, the generated authentication code is called an electronic signature because it is from a proprietary private key and no one else can falsify his/her signature. Reliability in e-Commerce and e-payment system is to prevent computer failure, procedural errors and transmission errors. Hardware failures, software errors, computer viruses and natural disasters can result to failure of e-payment systems and result to loss of reliability and consequent litigations [9].

IV. SYSTEM ANALYSIS OF RELATED FRAMEWORKS

The following section describes the system analysis of two types of related e-payment security frameworks or standards—the e-payment system without security and Payment Card Industry Data Security Standard (PCI DSS) adopted by the Nigeria apex Bank, the CBN.

A. E-payment Framework with no security

This electronic payment framework is assumed to have no form of security and therefore cannot be deployed in a mission-critical environment such as obtainable in Nigeria cashless electronic payment system where there are a lot security vulnerabilities and breaches on daily basis.

Shortcomings of this Epayment framework are as follows:

- 1) *It is the worst case because of lack of physical, operational and data security for all forms of epayment transactions. Transaction data, consumers' card data and personal information are not protected at all.*
- 2) *It allows hackers and fraudsters both within and outside the payment system with little or no effort to commit all sorts of cyber breaches and attacks and steal customers' funds.*
- 3) *It creates room for all forms of litigation since there will be much loss of privacy and loss of hard-earned funds.*
- 4) *It is not useful since it cannot be deployed in any sensible epayment system.*

B. PCI DSS

The Payment Card Industry Data Security Standard (PCI DSS) is an information security standard for organizations that handle cardholder information for the major debit, credit, prepaid, e-purse, ATM, and POS cards [10]. It was originated by the team from the card companies- MasterCard, VisaCard, PayPal and Microsoft in 2009 to address information security challenges arising from use of credit, debit and pre-paid cards on payment terminals such as ATM, PoS, Online.

Shortcomings of PCI DSS are as follows:

- 1) *It provides data security but very expensive to implement by an ordinary stakeholder such as merchant.*
- 2) *It is very complex and complicated to implement and not user-friendly. It therefore costs a lot of money and time to train IT staff to master it.*
- 3) *Because of fines and penalties imposed by the originators on defaulting stakeholders, a lot of suspicions are created on the real intentions for the introduction of the security framework thereby casting aspersions on the objectives of the framework.*

V. THE PROPOSED FRAMEWORK

In this section we describe a framework for enhancing the security for the cashless electronic payment system of Nigeria that simplifies the implementation and reduces the total cost of ownership (TOC) especially for the merchants and consumers (the majority of the epayment stakeholders). In Nigeria, the merchants and the consumers form the majority of the electronic payment stakeholders and in order to ensure full security compliance, these stakeholders must be accommodated; the total cost of ownership (TCO) must be reduced for the merchant while user education on security compliance must be transferred to the consumers at little or no cost to them. At the same time, complexity of implementation must be reduced and the security system made simpler in order to attract full compliance by all the stakeholders.

Our proposed security framework transfers the bulk of the burden of cost of compliance (COC) and cost of ownership (TCO) to the bigger and richer epayment stakeholders such as commercial banks, switching companies, card processing firms

and mobile money providers who will at the long run recoup their investments. Security vulnerability software can be provided by the switching company for other stakeholders to enable them carry out regular vulnerability system scan. The license for this software should be one-time life license updateable for a minimal fee per annum.

After thoroughly studying the current Nigerian epayment architecture (Figure 1) and the epayment models as depicted in Figures 2 to 6, we discovered that virtually all electronic payment transactions will require electronic payment switching, so the electronic payment switching companies such as Interswitch and ETransact have to bear bulk of the cost of providing data security to cover the merchants and consumers. Nevertheless, every stakeholder has a role to play to ensure full compliance of data, operational and physical security of epayment infrastructure and networks.

Our proposed framework suggests the following guidelines and standards for various categories of electronic payment stakeholders:

- *Security compliance guidelines for the consumers (on how to use credit and debit cards on payment terminals).*
- *Security compliance guidelines for the merchant (using Point-Of-Sale (PoS) and Online Ecommerce Portal).*
- *Security compliance guidelines for the Switching firms and card processing companies.*
- *Security compliance guidelines for the commercial banks.*
- *Security compliance guidelines for the Mobile Payment providers.*
- *Security compliance guidelines for the government/regulatory agency*

A. Security Compliance Guidelines for the Consumers

The consumer or customer is the least of the epayment stakeholders in terms of responsibilities of security compliance. His duties are merely to protect his credit/debit card pins from being exposed and stolen; also to report to his bank or card provider immediately or in the event of loss or misplacement of his credit/debit card so that the account will be blocked to protect the money in his account from being stolen.

Equally, the consumer should reconcile with his bank for every transaction via debit/credit alert confirmation in his mobile phone.

B. Security Compliance Guidelines for the Merchant stakeholder

The merchant deals directly with the consumers/customers. He sells goods or services to the consumer and the consumer pays him electronically via the epayment channels or terminals such as PoS, Mobile device, online eCommerce shop or portal as depicted in Figure 7. The primary concern of the merchant is for the consumer to pay him through any of the available epayment channels and his bank account successfully credited and he obtains a credit confirmation alert in his mobile phone from his bank before he can allow the consumer or customer to go away with his goods or services.

The security requirements for the merchant using the Point of Sale (PoS) channel are as follows:

- 1) To provide physical security for the PoS equipment so that robbers will not cart it away.
- 2) To install the PoS equipment in such a way as to prevent tampering by would-be-criminals either from outside or within his shop.
- 3) To capture the biometric feature of each customer (facial) using camera or close circuit television (CCTV) in case of fraud so as to identify the culprit by law enforcement agents.
- 4) To reconcile his account status and balance with his banker via credit alert confirmation to his mobile phone to ensure that his account is credited before allowing the consumer/customer to go away with the goods.

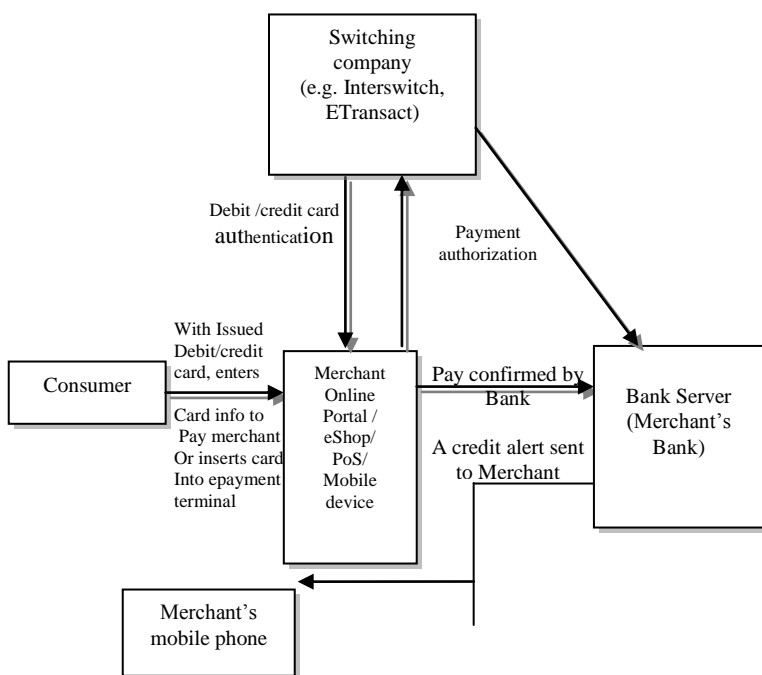


Figure 7. Merchant-customer epayment model using online eCommerce Portal, PoS and Mobile payment terminals

For the merchant-customer epayment model depicted in Figure 7, security vulnerability lies on the interfaces between the epayment terminals and the merchant's Bank Server. It is therefore the duties of the epayment switching companies to provide cryptographic data security through encryption and decryption and other additional data security such as digital signatures and hashing for every epayment transaction.

For the merchant-consumer model using online eCommerce shop or portal the merchant owes full responsibility to secure his online portal or shop to provide data security. The suggested guidelines for merchants operating online eCommerce portal or shops are listed as follows:

- 1) He has to obtain certificate authority(CA) to verify buyer/consumer's identities before selling,
- 2) He has to provide point-to-point encryption for every transaction through the provision of secure socket layer (SSL) on http interface (i.e. https), secure shell(SH), Internet Protocol Security(IPSec) etc to ensure data confidentiality and hashing function to ensure data integrity.

C. Security Compliance Guidelines for the Switching companies

The epayment switching companies play the major role in the entire electronic payment system in Nigeria. The switching firms conduct all debit/credit transaction switching from consumer credit/debit cards on epayment terminals such as ATM, PoS, mobile devices, online ecommerce portals to the authorized bank accounts of the merchants or electronic switching from a business account in a bank to another business account in another bank.

The channels or interfaces in the electronic switching, most of the time being wireless and wired Local (LAN) or Wide Area Networks (WAN) are vulnerable to all sorts of data and operational security attacks [11]. So the switching firms have to provide full compliance and protection for the data and operational capacity of all the epayment transactions and only charge other stakeholders minimal fees per annum.

The following security guidelines are suggested for the switching companies. They should:

- Provide end-to-end data encryption using strong cryptographic cipher engines (AES 128 bit) is suggested to ensure data confidentiality.
- Provide strong user authentication to provide data confidentiality and integrity.
- Provide Message Authentication Codes (MAC) or hashing functions to provide data integrity for all the epayment transactions.
- Perform vulnerability and compliance system scan on all interconnected stakeholders on their network on a regular basis (maybe quarterly) at a minimal fee to the stakeholders.

The channels or interfaces in the electronic switching, most of the time being wireless and wired Local (LAN) or Wide Area Networks (WAN) are vulnerable to all sorts of data and operational security attacks [11]. So the switching firms have to provide full compliance and protection for the data and operational capacity of all the epayment transactions and only charge other stakeholders minimal fees per annum.

D. Security Compliance Guidelines for the Commercial Banks

Commercial banks have a lot of customers that deposit their funds with them. They also deal directly with merchants who maintain business accounts with them. The epayment switching firms switch payment electronically from various epayment terminals such as ATM, PoS, Online Ecommerce portals, Mobile devices once the consumers authorizes payment using their credit/card cards.

Also switching firms switch payment electronically from one business account in one bank to another business account in another bank. All these epayment transaction switching end up in the banks' transaction servers which work 24 hours a day, 7 days a week. It is therefore the duties of both the commercial banks and switching firms to provide security to protect the Transaction server and the epayment interfaces from being security breached.

The following security compliance guidelines are suggested for the commercial banks to ensure compliance:

- Banks should provide physical security of ATM terminals to ensure availability of ATM terminal any time and to protect the ATM terminal from being attacked by armed robbers.
- Banks should provide end-to-end encryption of all transaction data from the TransactionServer to all epayment terminals via electronicpayment switching gateways. Powerful encryption engine such as AES (128 bit) is hereby suggested. This will provide data confidentiality and privacy.
- Banks should provide full authentication of their customers at ATM payment terminals using simple username and pin logins as well as additional biometric authentication mechanisms to capture customers' biometric feature such as photograph.
- Banks' core Banking solutions should be fully protected with hash functions or Message Authentication Codes (MAC) to protect the payment data from being compromised by fraudsters either from the within or outside. This will provide data integrity.
- Additionally, banks should ensure that all customers are sent instant SMS transaction alert or confirmation through their mobile phones to enable account reconciliation.

The following flowchart (Figure 8) depicts the proposed enhanced modified security framework for the protection of transaction data from a consumer (paying with his debit/credit card) to merchant account in the bank using electronic switching platforms.

E. Security Compliance Guidelines for the government/regulatory agency

The regulatory agency, the Central Bank of Nigeria (CBN) can spearhead the passing into law security compliance and liability bill that will force every epayment stakeholder to become compliance and to accept responsibility in the event of any financial loss to a business or a consumer as a result of lack of security compliance. The legislative bill should equally spell out who should pay who in the event of security liability. Also penalties should be clearly spelled out for would-be-criminals and fraudsters using long jail term sentences to discourage potential financial criminals.

CBN should also educate the users (consumers) and the other stakeholders on the need to be security compliance and how to be.

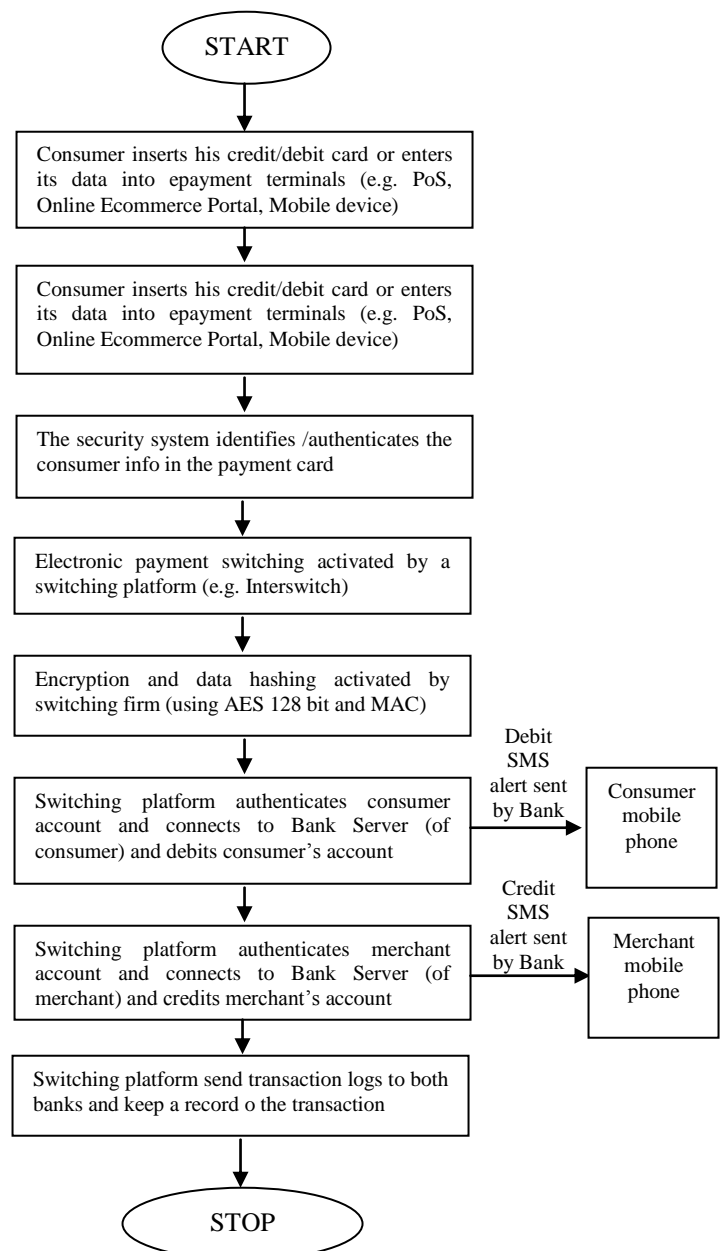


Figure 8. System flowchart depicting the proposed enhanced modified data security framework for Nigeria cashless epayment system from consumer to merchant accounts in the banks

VI. CONCLUSION

In this paper, the myriads of security as well technical and legal challenges facing the successful transition from cash-based to cashless electronic payment of Nigeria epayment system have been identified. The investigations have shown that the security models of an epayment system must be studied and understood before a data security framework is adopted for it. This study have revealed why the CBN adopted data security framework, the PCI DSS, failed to attract wide acceptability and compliance in Nigeria's epayment system. Issues of complexity in implementation and high prohibitive cost of implementation as well as lack of legal framework have been identified to be the major bottlenecks towards the compliance to the PCI DSS by major epayment stakeholders in Nigeria.

The result of the findings from this study shows that for full compliance to any adopted data security framework for any epayment system, cost and simplicity of implementations must be seriously considered.

In this paper, we introduced a simpler and less costly epayment security framework/standards which we believe will provide enhanced data and operational security as well reduce cost and complexity of implementation by the majority of the stakeholders in Nigeria cashless economy. By reducing the complexity and cost of implementation as identified in PCI DSS, the majority of the epayment stakeholders such as consumers and merchants will easily and willingly become security compliant.

The Nigerian apex bank, CBN, should seriously look inwards toward encouraging the major stakeholders such as consumers and merchants to comply with data security guidelines by adopting framework that will lower their overall cost of compliance (COC) and total cost of ownership (TCO) and as well reduce complexity of implementations considerably. It is in this regard we sincerely hope that our contributions will help in no small measure in providing the roadmap for enhanced, simpler and cheaper security standards and framework for the new Nigeria cashless electronic payment system.

REFERENCES

- [1] Central Bank of Nigeria, "Further Clarifications on Cashless Lagos Project". Retrieved on 9th September, 2012 from <http://www.cenbank.org/cashless>.
- [2] Mynaij.com. "Cashless Banking: Cyber Criminals Now Focus on Nigeria". Retrieved on 6th August, 2012 from http://news.naij.com/business_and_economy/.
- [3] Mynaij.com. "Cashless Banking: Cyber Criminals Now Focus on Nigeria". Retrieved on 8th August, 2012 from http://news.naij.com/cashless_epayment_in_Nigeria/.
- [4] Businessdayonline. "Cyber Security and Nigeria's Cashless Initiative", Retrieved on 9th September, 2012 from <http://www.businessdayonline.com/NG/index.php/analysis/editorial/4329/cyber-security-and-nigerias-cashless-initiative>.
- [5] Wikipedia, "Payment Card Industry Data Security Standard". Retrieved on 9th September, 2012 from <http://www.wikipedia.com/wiki/Payment-Card-Industry-Security-Standards.htm>.
- [6] S.G.P. Muniappa and S.A.Gosh, "Report on Internet Banking: A White paper on Indian Internet Banking", p. 12-14.

- [7] C.K. Ayo and W.I. Ukpere, "Design of a secure unified epayment system in Nigeria: A Case Study", African Journal of Business Management, vol.4 (9), pp. 1753, August 2010.
- [8] Y. Jing, "On-line Payment and Security of Ecommerce", Proceedings of the 2009 International Symposium on Web Information Systems and Applications (WISA '09) China, May 22-24, 2009, pp. 546, 2009.
- [9] Dan Zhu, "Security Control in Inter-Bank Fund Transfer", Journal of Electronic Commerce Research, vol.3, No.1, pp. 46-48, 2002.
- [10] Wikipedia, "Payment Card Industry Data Security Standard". Retrieved on 18th August, 2012 from <http://www.wikipedia.com/wiki/Payment-Card-Industry-Data-Security-Standard.htm>.
- [11] PCI Security Standards Council, "Protecting Telephone-based Payment Card data", pp.1.

AUTHORS PROFILE

Engr. Fidelis Chukwujekwu Obodoze is a Ph.D research scholar in the Department of Electronic and Computer Engineering NnamdiAzikiwe University Awka, Nigeria. He is currently lecturing at the Department of Computer Science Renaissance University Enugu, Nigeria. His research interests include Cryptography, Wireless Telecommunication security, Wireless Sensor Networks, Information and Data security, Artificial Intelligence and Industrial Automation. He had his Masters (MEngr.) in Control Systems and Computer Engineering at Nnamdi Azikiwe University Awka, Nigeria in 2010 and Bachelor of Science degree (B.Sc.) in Computer Engineering at ObafemiAwolowo University Ile-Ife, Nigeria in 2000. All correspondence to email: fidelisobodoze@gmail.com

Dr. Francis A. Okoye is a lecturer and Head Department of Computer Science and Engineering Enugu State University of Science and technology (ESUT), Nigeria. He had his Ph.D in Computer Science in 2008 at Ebonyi State University Abakaliki; MSc. and BEng. in Computer Science and Engineering at Enugu State University of Science and Technology (ESUT), Nigeria in 2001 and 1996 respectively. Email: franciscd@esut.edu.ng

Mr. Samuel Chibuzor Asogwa is a Ph.D research scholar in the Department of Computer Science Nnamdi Azikiwe University Awka, Nigeria. He is currently lecturing at the Department of Computer Science Michael Okpara University of Agriculture Umudike, Nigeria. He had his Masters (MSc.) in Computer Science at Ebonyi State University Abakaliki, Nigeria in 2011 and Bachelor of Engineering (BEng.) in Computer Science and Engineering at Enugu State University of Science and Technology (ESUT), Nigeria in 1999. Email: sasogwa@gmail.com

Mr. Frank EkeneOzioko is the Director of ICT Unit and Lecturer Department of Computer and Information Science, Enugu State University of Science and Technology (ESUT), Nigeria. Email: ekene.oziko@esut.edu.ng

Engr. Calista Nnenna Mbah is a Lecturer and former Head, Department of Computer Engineering, Caritas University Enugu, Nigeria. She is currently pursuing her Ph.D programme in Computer Science at Nnamdi Azikiwe University Awka, Nigeria. She had her MSc. Computer Science at University of Nigeria Nsukka (UNN) in 2009, Nigeria and B.Eng in Computer Science and Engineering at Enugu State University of Science and Technology (ESUT), Nigeria in 2004. Email: mbacally@yahoo.com

Cashless Society: An Implication To Economic Growth & Development In Nigeria

¹N. A. YEKINI, I. K. OYEYINKA², O.O. AYANNUGA³, *O. N. LAWAL⁴, and A. K. AKINWOLE⁵

^{1,2,3,4,5} Department of Computer Technology, Yaba College of Technology, Yaba, Lagos Nigeria

Abstract—The Central Bank of Nigeria (CBN) announced a new cash policy (Cashless Society), which began on 1st of January 2012 in Lagos state. The policy will be effective nationwide from June 1st 2012. The various technologies and issues involved in cashless society are discussed. This research work sampled the opinions of the general public on issues surrounding the Cashless Society by using structured questionnaire to gather information from people. Data obtained was tabulated and the result was analyzed using percentages. The analyzed result was presented graphically and it was discovered that larger percentage of sampled population are of the opinion that Cashless Society will proffer solution to high risk of cash related crimes such as illegal immigration, financing Terrorism, illegal drug Trade, Human Trafficking, corruption etc. It will enhance effectiveness of monetary policy in managing inflation in Nigeria, thereby improving economic growth and development. Various issues such as benefits and social implication of Cashless Society are discussed. We proposed a Model Cashless System for Nigeria, which will be more convenient for people to appreciate, embrace, and use.

Keywords- Cashless Society; Corruption; Development; Economic growth; Electronic Card.

I. INTRODUCTION

The CBN (2011) introduced a new policy on cash-based transactions which stipulates a ‘cash handling charge’ on daily cash withdrawals or cash deposits that exceed ₦500,000 for Individuals and ₦3,000,000 for Corporate bodies. The new policy on cash-based transactions (withdrawals & deposits) in banks, aims at reducing (NOT ELIMINATING) the amount of physical cash (coins and notes) circulating in the economy, and encouraging more electronic-based transactions (payments for goods, services, transfers, etc.). The new cash policy was introduced for a number of key reasons, including:

1. To drive development and modernization of our payment system in line with Nigeria’s vision 2020 goal of being amongst the top 20 economies by the year 2020. An efficient and modern payment system is positively correlated with economic development, and is a key enabler for economic growth.
2. To reduce the cost of banking services (including cost of credit) and drive financial inclusion by providing more efficient transaction options and greater reach.
3. To improve the effectiveness of monetary policy in managing inflation and driving economic growth.

In addition, the cash policy aims to curb some of the negative consequences associated with the high usage of physical cash in the economy, including:

- **High cost of cash:** There is a high cost of cash along the value chain - from the CBN and the banks, to corporations and traders; everyone bears the high costs associated with volume cash handling.
- **High risk of using cash:** Cash encourages robberies and other cash-related crimes. It also can lead to financial loss in the case of fire and flooding incidents.
- **High subsidy:** CBN analysis showed that only 10% of daily banking transactions are above ₦150,000, but the 10% account for majority of the high value transactions. This suggests that the entire banking population subsidizes the costs that the tiny minority 10% incur in terms of high cash usage.
- **Informal Economy:** High cash usage results in a lot of money outside the formal economy, thus limiting the effectiveness of monetary policy in managing inflation and encouraging economic growth.
- **Inefficiency & Corruption:** High cash usage enables corruption, leakages and money laundering, amongst other cash-related fraudulent activities.

The Content of the Cash policy shall apply from January 1st 2012 in Lagos State (“tagged Cash-less Lagos”) as follows:

- Only CIT licensed companies shall be allowed to provide cash pick-up services. Banks will cease cash in transit lodgment services rendered to merchant-customers in Lagos State from December 31st 2011. Any Bank that continues to offer cash in transit lodgment services to merchants shall be sanctioned.
- 3rd party cheques above ₦150,000 shall not be eligible for encashment over the counter. Value for such cheques shall be received through the clearing house.

The idea is, among others, to identify the habits and preferences of the users, be it a teenager without a credit/debit card or a retired elderly with a checkbook, and make the mobile phone the new electronic wallet and thereby try to abolish physical money: cash. Furthermore, the goal is to measure the experience when paying without physical money and compare it to an ordinary physical payment in terms of safety, complexity, efficiency, for example (Hansen, 2011). The Cashless Policy (aka Cashless Society) of Central Bank of Nigeria has begun on the 1st of January 2012 in Lagos State, Nigeria. This policy will be effective nationwide from

June 1st 2012. Several apprehensions have been expressed by the public on this policy but the Central Bank of Nigeria has educated the general populace on the need to embrace this policy as it will be useful in solving cash related crimes as listed above, enable monetary policy to tackle inflation in the economy, reduce cost of banking services like reduction in cost of printing money and transportation of cash from one location to another.

This Research work sampled the opinions of the general public on issues surrounding the Cashless Society policy and its implication on economic growth and development in Nigeria. Several countries in Europe, Asia and other continents have been operating cashless society for a long time and it has been beneficial to their respective economies. It solved cash related crimes in those societies like financing terrorism, kidnapping, money laundry etc.

Section 2 of this paper discusses some of the various technologies related to Cashless society. Section 3 presents research methodology, which was based on sampled population of nine thousand and eighty (9,080) individuals that cut across students above eighteen years, civil servants, market men and women, and professional (including bankers). Information was gathered using questionnaire. The data obtained was analysed and discussed. Some issues emanating from the discussion were presented in section 4. We propose a Cashless Society Model for Nigeria in Section 5, which we hope should be convenient for people to embrace and use. Section 6 concludes the paper, while recommendations are presented in Section 7.

II. LITERATURE SURVEY

A. Electronic Cash

Several companies have taken this idea further and developed cards which can be used in multiple retail outlets, effectively as “electronic cash”. One such system is Mondex, developed by the National Westminster Bank in the UK and later sold to MasterCard International. Mondex was originally developed in 1996 as a “smartcard alternative to cash”. Graham Higgins, a banker and co-inventor of Mondex, had been quoted as explaining that the scheme would help alleviate “the burden of counting, storing, as well as the security associated with physical cash” (Bonugli, 2006).

B. The Advent Of Plastic Cards

Card-based alternatives to cash payments are now well established, with credit and debit cards in popular usage. Additionally, new technology has enabled the development of so-called ‘smartcards’ where additional data can be stored on a microchip Chip and PIN Press Office 2005, 2006).

1) Smart Cards

A smart card is a plastic card, similar in appearance to a credit card, and containing one or more embedded semiconductor chips. Smart cards typically have a storage area in EEPROM and may also include a microprocessor able to process any data stored. Recent technological progress has seen the development of a “contactless” smart card, one in which the chip communicates with a card reader using radio frequency identification (RFID). Smart cards have significant

potential over magnetic-stripe ‘swipe’ cards: not only can more data be stored, but it can be processed in some way as well. Despite privacy concerns, it seems likely that smart cards are the way forward, with increasing systems merging together. In an article for Credit Union Magazine, Schacklett (2000) predicts that “as smart cards gain momentum in the financial services marketplace, it is likely that other forms of plastic like credit, debit, and ATM cards will all melt into one universal, multifunctional smart card”.

The first major use of smartcards was by French banking association, CartesBancaires, which saw advantage of using the technology in reducing fraud. By replacing magnetic striped cards with smart cards; fraud rates in France dropped tenfold (Flohr, 1998).

2) Credit And Debit Cards

In the 1950s, Diners’ Club began issuing charge cards and since then, a major proliferation in credit card usage has made such cards become a popular alternative to cash payments. In the thirty years between 1971 and 2001, the number of cards per household in the United States grew from 0.8 to 7.6 (PayingWithPlastic.org, 1997).

3) Stored-Value Cards

Stored value cards are typically similar in appearance to credit cards and either employs a magnetic stripe or smart card technologies in order to store data. Under this scheme, using an appropriate reader an amount can be electronically added or deducted from a balance on the card. Such a scheme is seen by some as an “initial step toward a cashless society” (Shelfer and Procaccino, 2002).

C. Security Issues In Cashless Society

Security is clearly of crucial importance in considering any alternative to physical cash. At the root of this lies the problem of authentication, “the process of verifying the identity of a person” (Pountain, 2003). This is typically performed by examining some identifying information such as a password or digital signature.

1) Pins

One of the obvious and most commonly used forms of authentication is a password. In the context of payment systems more commonly implemented as a personal identification number (PIN). Such a system has long been in place for authenticating users of cash points prior to withdrawing money. In line with much of Europe, since 2003 the UK has been in the process of converting to use of a similar system for the authentication of credit or debit card transactions. Since the 1970s, most face-to-face card transactions have made use of a magnetic stripe card to read and record account data, along with the customer’s signature for verification. However, technological advances meant that criminals have been increasingly successful in making copies of the data stored on the magnetic stripe, and forging signatures in order to commit fraud. Over £402 million was lost through “plastic card fraud” in 2003, which has led to the advancement of a new system, marketed in the UK under the name ‘Chip and PIN’ (British Retail Consortium, 2005).

This system sees a replacement of magnetic stripe cards, with ‘smart cards’ containing an integrated-circuit chip which

can store and verify a PIN. Rather than sign for purchases in a retail outlet, customers are required to input their PIN in a reader (Chip and PIN Press Office, 2005). Uptake of the 'Chip and PIN' system has, on the whole, been very responsive and it is planned that signatures will cease to be accepted as verification of a chip card from February 14th 2006. Official statistics show that by the end of 2005 more than 80% of shops and card accepting businesses had upgraded their point of sale equipment to accept 'Chip and PIN' and that 99% of cardholders in the UK had at least one 'Chip and PIN' enabled card (Chip and PIN Press Office, 2006).

The 'Chip and PIN' system is not without criticism, however. One of the main problems is that the introduction of PIN based systems in securing credit card payments is that they will not protect 'card not present' transactions (such as those placed over the internet or by telephone). Thus, there is still a market for stealing or counterfeiting credit cards which can still be used without knowledge of the PIN (Anderson, Bond, and Murdoch (2010).

2) Biometrics

Biometrics is "an area of information technology concerned with the study of techniques capable of uniquely identifying a human individual, based upon intrinsic physical or behavioral traits" (Wikipedia, 2011). In the realm of security, particularly in terms of electronic payments, these characteristics may then be adopted for authentication purposes.

III. RESEARCH METHODOLOGY

We sampled the opinion of nine thousand and eighty (9,080) individuals within Lagos state. The sampled population cut across students above 18 years, civil servants, market men and women (both educated and illiterate), and professionals

(including bankers). The medium of communication is English language, but people were employed to administer the questionnaires. For those who were unable to read and write in English language, the questionnaire was translated to them in their native language and the person that administered the questionnaire then tick the answer based on their response. Data was gathered from the sampled population by using structured questionnaire with easy means of response, which is based on ticking the best option from the given available options: YES, NO, and UNDECIDED.

The responses to the questionnaire were recorded and tabulated raw as in Table 1, and in percentages as in Table 2 below:

Table 1: Raw Data Obtained from Questionnaire

Questions#	Response		
	Yes	No	Undecided
1	4994	4086	0
2	5902	3178	0
3	5448	2724	908
4	1543	7264	273
5	3178	5448	454
6	6447	1906	727
7	5539	1998	1543
8	4630	1725	2725
9	5357	2815	908
10	1907	6447	726

Table 2: Data Obtained from Questionnaire in Percentage

Questions#	Response		
	Yes (%)	No (%)	Undecided (%)
1	55.00	45.00	0.00
2	65.00	35.00	0.00
3	60.00	30.00	10.00
4	16.99	80.00	3.01
5	35.00	60.00	5.00
6	71.00	20.99	8.01
7	61.00	22.00	16.99
8	50.99	19.00	30.01
9	59.00	31.00	10.00
10	21.00	71.00	8.00

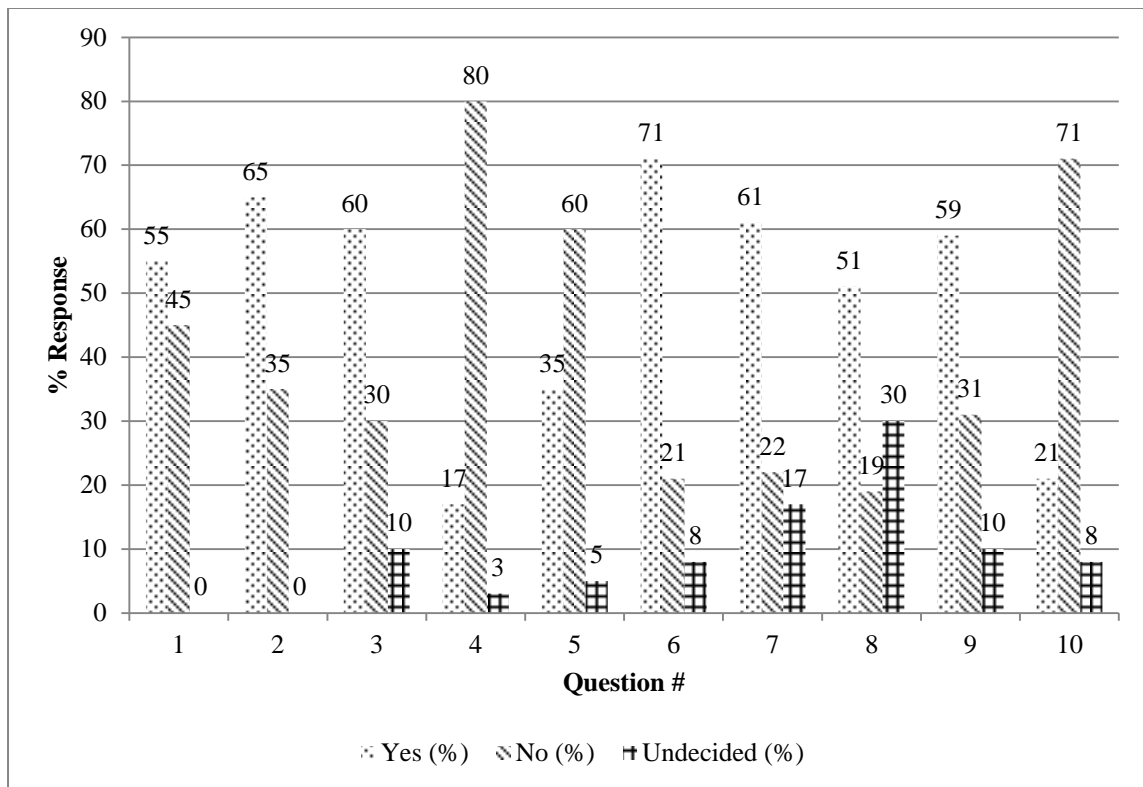


Figure 1: Chart showing response % from Table 2

The result is discussed in section 4.1.

IV. DISCUSSION/CRITICAL ANALYSIS OF THE RESULT

4.0.1 General Discussion On Response To Questionnaire

Question 1: Are you aware of the cashless banking policy recently announced by Central Bank of Nigeria (CBN)?

55% and 45% of the people which are equivalent to 4,994 and 4,086 were respectively aware and unaware of the policy. This close gap indicates that the stakeholders in this policy should intensify effort in creating awareness among the general public on issues regarding cashless policy.

Question 2: Are you currently operating a bank account with any bank in Nigeria or outside Nigeria?

65% of the sampled population has bank account while 35% does not. In this case, those 35% (3178) cannot fully participate in cashless society. This 35% were among the students, and market women. Consequently efforts must be put in place to encourage these individuals to open bank account so that they can be fully part of cashless banking policy.

Question 3: Is this Cashless Policy (Cashless society) a welcome idea to you as an individual?

60% welcomed the cashless banking policy while 30% does not, and 10% undecided. If enough awareness and education is given to the general public, we strongly believed that more people will be willing to support the policy rather than forcing on the people.

Question 4: Do you think Nigeria has enough infrastructures such as electricity, IT Personnel etc. required for successful implementation of the policy?

17% believed that Nigeria has enough infrastructure and trained personnel to support the cashless policy. 80% said No to this question and 3% were undecided. There is no doubt that infrastructures to support this policy such as stable electricity, system networking etc. are seriously unavailable for full implementation of the policy. Therefore it is advisable that government starts implementation of this good policy in phases and come out with model that will be more convenient for people to embrace.

Question 5: Do you think the government and the major stakeholder in the banking policy in Nigeria have given the proper information, awareness or education on the need for the adoption of a cashless banking policy in Nigeria?

Majority of the people believed that people were not well informed on the issues surrounding this policy as 60% said No, and 35% said Yes, while 5% were undecided. Therefore people need to be more educated.

Question 6: Do you think Cashless banking will help to reduce high-risk of cash related crime in Nigeria?

71% said Yes, 22% said No and 7% were undecided. There is overwhelming consensus that cashless policy will be a tool to eliminate or reduce to miniature the cash related crime such as: illegal immigration, financing Terrorism, illegal drug Trade, Human Trafficking, corruption etc. in Nigeria.

Question 7: Do you think high usage of cash in the society is one of the reasons for the high cost of banking services in Nigeria?

61% agreed that high usage of cash is one of the reasons for high cost of banking services in Nigeria, while 22% said No and 16.9% were undecided. In this case it was discovered that cashless banking will reduce the cost of systems as cost of printing, and transportation of cash will be reduced considerably.

Question 8: Cashless banking will improve the effectiveness of the monetary policy to managing inflation?

51% said Yes, 19% said No, and approximately 30% were undecided. It was discovered that Yes response were among educated people that know the technicality and principles involved in using monetary policy to tackle inflation.

Question 9: Cashless Society will reduce high cost of printing and transporting cash from one financial institution to another?

59% said yes 31% said no and 10% were undecided. In this case it was concluded that the policy will surely reduce the high cost of using and printing cash as well as nuisance associated with transportation of currency from one location to another.

Question 10: Do you think government and the stakeholders in this policy have done enough in term of security of public funds that may arise from cashless banking?

71% believed that government have not done enough in area of security, 21% response in contrary an approximately 8% were undecided. Security is clearly crucial and important consideration of any alternative to physical cash. In this era of cashless society different form of card (smartcard) will be introduced. At the root of this lies the problem of authentication, which results from using those cards? In this case PIN (personal Identification Number), should be replaced with Biometric features such as finger prints, facial recognition, voice recognition etc., for robust security. Based on the critical analysis and general discussion of the result of responses of sampled population, we hereby pinpoint the Benefit of Cashless Society in section 4.1, Social implications of Cashless Society in section 4.2, and our Proposed Cashless Society Model for sustainable cashless society in Nigeria in Section 5.

A. Benefits Of Cashless Society

There is no doubt that with cashless society, the use of paper and coins currencies will reduce, migrating to virtual

money that uses computerized systems to handle transactions. Replacing physical cash with cashless credits or electronic money transfer will help to:

- Eliminate or reduce cash related crimes such as illegal immigration, financing terrorism, illegal drug trade, human trafficking, money laundering, corruption etc. in Nigeria.
- A Cashless Society policy will go a long way in making our society a better place to live with a reduced rate of criminal activities.
- Paper cash is non-traceable and unaccountable; it can be easily hid/lost, stolen, counterfeited, and spent without a trace. As a result physical cash has allowed all sorts of criminal activities to thrive. However in a cashless economy, this will change overnight. Certain crimes like armor car heists, kidnap for ransom, store holdups, armed robberies, will cease since there will be no paper cash to steal.
- Reduce corruption.
- Reduce cost of printing, managing and transporting paper money, thereby making more funds available to the government, which will translate to economic growth and development.
- Facilitate possibility to carry large quantities of money around safely and effectively, even an entire personal wealth.

B. Social Implications Of Cashless Society

This cashless society also has its cons, some of these include:

- Government would be able to monitor purchase, spending habits and businesses patronized. In this case government will have a total control over everything we do. It can oppress or suppress society where it hurts the most and that is money.
- In case of financial crisis, how could we backup the purely electronic financial reserves?
- Business will be conducted with imaginary money; this may reduce the value of money and bargaining power.

V. PROPOSED CASHLESS SOCIETY MODEL

Figure 2 is the diagram for our proposed “Cashless SocietyModel”.

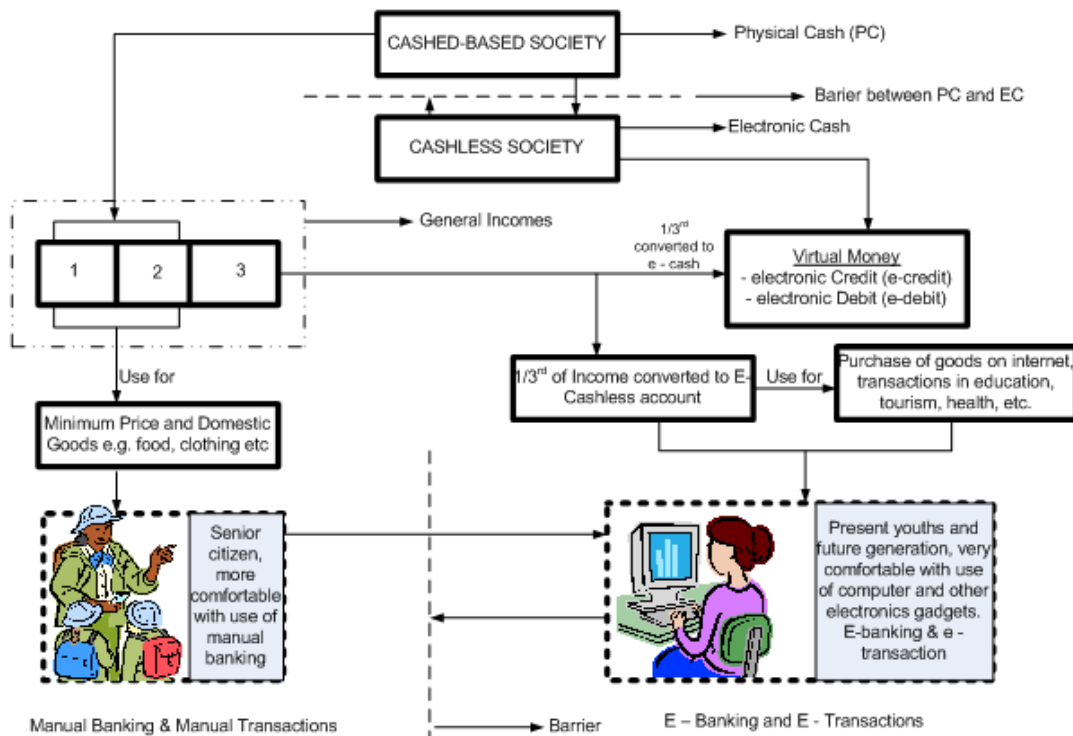


Figure 2: Proposed Cashless Society Model

Figure 2 is discussed as follows: the cashless society will run parallel to the ongoing cash-based society. The model presented will work as itemized below:

- Electronic cash will exist as E-credits and e-debits. The e-credits and e-debit will function as the currency or cash as in present times.
- 1/3 of a person's general income will be converted to e-cash, so that a part of the person becomes a cashless.
- Cashless e-credit policy shall be used in the following areas of transactions:
 - Purchase of any types of electronic goods
 - Purchase of machines and hardware components
 - Transaction in education and health care systems
 - Transaction in travel and tourism (e-booking).
 - And in many other similar transactions.
- The cashless transactions should not be used in transaction dealing with essential commodities such as foods, clothing etc. in the meantime.
- The cashless society must be mainly planned for the young and future generation of people. Senior citizens (i.e. the elderly people) could be left out. But doors should be open for whoever among the senior citizens that is interested to cross the barrier and move to cashless transactions.
- Cashless credit should not be converted to real cash, but cash based money can be converted to cashless.
- E-banks should be available to manage and maintain cashless account; these banks will just be website and called cashless banks and will maintain and manage the credits, debits and savings and record the transactions

made by every account holders. This will eliminate the manual banking system which is relevant today.

With this hypothetical concept for a cashless society, it would be beneficial for the present and future generations. This concept will enable Nigerian economy to be digitized and provide a unique convenient system of cashless transactions. It will also enable step by step migration to cashless society.

VI. CONCLUSION

A cashless Society is a welcome idea in this modern day, because the use of credit and debit cards has become an established and popular alternative to cash. It is then advisable for every individual to embrace this policy. Some important issues have been identified and some will be encountered as people embrace this policy. Some of these issues include transfer to electronic transaction, benefits and social implication, security etc. This paper discussed some of the benefits of cashless society towards economic growth and development, especially in the area of eliminating corruption and reducing the cost of producing, transporting and managing cash money. A Cashless Society Model was proposed which, if implemented in a proper manner, can engender a perfect cashless transaction system in future, which will improve our economic growth and development and reduce corruption among others.

VII. RECOMMENDATIONS

- It is highly recommended that this research should be carried out in other states of the federal republic of Nigeria before full implementation of the cashless society in June 2012. It is not possible to discuss all of the issues and investigation during compilation of this paper.

- Government should create more awareness and intensify efforts in educating the general public on issues related to cashless society, improve on infrastructure and training of personnel so that cashless society will be more acceptable, convenient and durable in Nigeria.
- Government should intensify effort to improve the state of electricity supply.
- Government should harness the use of our satellites in space to improve the speed of computer network to true broadband.
- The citizens should brace up to understand the positive impact of the policy, so that they can embrace it, for their own good.

REFERENCES

- [1] British Retail Consortium 2005. *BRC Yearbook 2005*. The Stationery Office, London, UK.
- [2] CBN 2011. *Further Clarifications on Cash-less Lagos Project*. Central Bank of Nigeria, Abuja. Retrieved from <http://www.cenbank.org/cashless/> 6th April, 2012.
- [3] Chip and PIN Press Office 2005. *EMV - Chip and PIN Technology*. Press release, Chip and PIN Press Office, London, UK.
- [4] Chip and PIN Press Office 2006. *I~PIN*. Press release, Chip and PIN Press Office, London, UK.
- [5] Hansen S. D. 2011. *The Cashless Society and Its Challenges*. Aalborg University, Denmark.
- [6] Bonugli P. K. 2006. *The Cashless Society: Increased Usage of Card-Based Payment Systems*. School of Electronics and Computer Science, University of Southampton, United Kingdom.
- [7] PayingWithPlastic.org 1997. *Household usage: the past thirty years*. <http://www.paywithplastic.org/index.cfm?gesture=statsDetailPrinter&aid%=1320>.
- [8] Pountain D. (ed) 2003. *Concise Dictionary of Computing*. Penguin Reference, London, UK.
- [9] R. Anderson, M. Bond, and S.J. Murdoch 2010. *Chip and Pin is Broken*. Technical report, Computer Laboratory, University of Cambridge, UK.
- [10] Schacklett M. 2000. *These business trends will shape the future of e-commerce*. Credit Union Magazine.
- [11] Shelfer K. M. and Procaccino J. D. 2002. Smart Card Evolution. *In Communications of the ACM*, 45:83–88.
- [12] Flohr U. 1998. The smartcard invasion. Byte.Wikipedia 2011. Biometrics. From <http://en.wikipedia.org/wiki/Biometrics>

The Analysis of the Components of Project-Based Learning on Social Network

A Case Study of Presentation Skill Course

Kuntida Thamwipat

Department of Educational Communications and
Technology,
King Mongkut's University of Technology Thonburi,
Bangkok, Thailand

Napassawan Yookong

Department of Educational Communications and
Technology,
King Mongkut's University of Technology Thonburi,
Bangkok, Thailand

Abstract—This research aims to analyze the components of project based learning on social network, case study of presentation skill course. The study was carried out with 98 representatives of 2nd – 3rd year students of Educational Communications and Technology Department, 2nd semester, academic year 2554, who was studying ETM 205 Presentation Skill. 5 point-rating scale questionnaire was used as the research tool for data collection for analyzing the obtained data which has the reliability as 0.94 and using the mean, Standard Deviation (S.D.), factor analysis, varimax method in data analysis. The results can be concluded as follows: Ten components of project based learning on social network, case study of Presentation Skill course was divided into 3 steps following Ministry of Education's framework as follows: 1) Pre-Project Stage consists of the 3rd component: The introduction of presentation skill and sharing ideas through social network, the 5th component: The learner preparation of pre-project stage, the 9th component: Writing the schemes and skills development training and the 10th component: The presentation skill training in group and individual 2) While / During-Project Stage consists of The 2nd component: Roles and responsibilities during project stage and communication, the 6th component: Readiness of performing duties of individual or group of people which communicate through social network and the 8th component: Guidance and agreement which learner submit voluntarily 3) Post-Project Stage consists of the 1st component: Evaluations, document management system, presentation, the 4th component: Instructor evaluation, feedback, congratulations to learners who achieve successful project and the 7th component: Multi-channel evaluation by regression equation or forecast equation concerning the analysis of the components of project learning management on social network, results as below:

$$Y = .780(\text{Factor } 1) + .603(\text{Factor } 2) + .686(\text{Factor } 3) + .591(\text{Factor } 4) + .741(\text{Factor } 5) + .912(\text{Factor } 6) + .632(\text{Factor } 7) + .824(\text{Factor } 8) + .689(\text{Factor } 9) + .686(\text{Factor } 10)$$

This forecast equation has 60 percent accuracy and deviation as 10.00

Keywords-analysis of the componenst; project-based learning; social network; Presentation Skill course.

I. BACKGROUND AND RATIONALE

Thai society at the moment has changed dramatically and rapidly in terms of economic, political, social and cultural aspects. The consequence from such changes means the development of sciences which grows in a rapid manner. According to numerous studies, computer has played a major role in such changes and it is widely used. Since computer has been upgraded all the time, users can have more choices to meet their demands and requirements. At the present moment, internet is becoming more popular and more people are surfing the internet for communication, information exchange, knowledge and entertainment to keep informed of the current situations.

Social networks are beneficial because they enable people to connect, to make friends, to communicate and to provide services to many groups of people. These websites usually contain bases for people to come and make friends through their tools to facilitate their online community. Users can establish their groups of the same interest and get connected to those who they are familiar with. The tools which they use could be e-mail, web board, and blog. Every user can share their profile, information and interesting facts. Besides contact with their friends, users can contact the friends of their friend as well [1].

Nowadays, internet and World Wide Web provides more social network websites. It could be said that social network can create the relationship between you and your friend through the circle of friends. The examples of this kind of social network websites are hi5, facebook and twitter.

Therefore, the use of social network websites for the instruction in classrooms is adopted as a supplement after class in order to give more opportunities of conversation and debate between instructors and students. Now social network is considered to be one approach in education [2].

Project-Based Learning allows learners to interact with communication and real life experience in terms of contents and language for communication during their project. Learners have opportunities and choices to make and they will respect others' opinions.

Moreover, they need to motivate themselves to gain authentic interest in the subject matter. With Project-Based Learning, learners encounter systematic working in accordance with plans, work as part of the team, exercise their creativity, build up their confidence, and more importantly, create a new body of knowledge on their own. Still, learners need guidance and supervision and this is the gap in which social network can prove useful because instructors could use social network for communication with students.

Presentation Skill course is a course from Department of Educational Communications and Technology, Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi. In this research case study, it was held in the second semester. Instructor in this course applied Project-Based Learning for instruction along with various activities for students to participate. Therefore, the researchers needed to analyze the components of Project-Based Learning on social network with a case study on Presentation Skill course in order to know the proper components for the design of Project-Based Learning on social network in the future.

II. RESEARCH OBJECTIVE

This research on the analysis of the components of Project-Based Learning on social network with a case study on Presentation Skill course was aimed to analyze the components of Project-Based Learning on social network and the Presentation Skill course was used as a case study.

III. EXPECTED OUTCOMES

1. Instructors would know the proper components of Project-Based Learning on social network and could use these results to develop and revise the current instructional conditions of Project-Based Learning on social network.

2. This research would become useful for the development of Project-Based Learning on social network in the future.

IV. RESEARCH SCOPE

A. Population in This Study

The population in this study comprised 98 students in their second or third year in the second semester of the academic year 2554 B.E. (2011). They were from the Department of Educational Communications and Technology, King Mongkut's University of Technology Thonburi who registered in ETM 205 Presentation Skill course [3].

B. Tool Used in This Study

The questionnaire was used in this study. Its aim was to analyze the components of Project-Based Learning on social network with a case study on Presentation Skill course. There were many items for each component of Project-Based Learning on social network. The questionnaire was based on 5-point rating scale.

C. Experts

There would be 3 experts in educational instruction and presentation with at least a Master's degree or 5 years of experience in educational instruction and presentation.

V. TOOL USED IN THIS RESEARCH

The tool used in this research was the questionnaire on items which affect Project-Based Learning on social network.

The developed questionnaire was designed to analyze according to the frequency and it was based on Likert's 5-point rating scale. The data would be analyzed using mean and standard deviation.

The statistical approach to this research was factor analysis with orthogonal factor rotation, i.e. Varimax method.

VI. RESEARCH RESULTS

Table I shows the confirmatory factor analysis of 10 factors.

TABLE I. CONFIRMATORY FACTOR ANALYSIS OF 10 COMPONENTS

Rank	Component	Weight
1	Readiness of performing duties of individual or group of people which communicate through social network	.912
2	Guidance and agreement which learner submit voluntarily	.824
3	Evaluations, document management system, presentation	.780
4	The learner preparation of pre-project stage	.741
5	Writing the schemes and skills development training	.689
6	The introduction of presentation skill and sharing ideas through social network	.686
7	The presentation skill training in group and individual	.686
8	Multi-channel evaluation	.632
9	Roles and responsibilities during project stage and communication	.603
10	Instructor evaluation, feedback, congratulations to learners who achieve successful project	.591

a. Standard error of estimate was 10.00

b. Cumulative predictive power was 60.00 %

According to the data analysis, it was found that 10 independent variables or components have a correlation with Project-Based Learning on social network with a case study on Presentation Skill course. The value ranged from .591 to .912 and the correlation was high. All 10 components formed cumulative predictive power of 60.00 % and the standard error of estimate was 10.00. The weight of components which affect Project-Based Learning on social network could be rearranged as follows: 1) Readiness of performing duties of individual or group of people which communicate through social network; 2) Guidance and agreement which learner submit voluntarily; 3) Evaluations, document management system, presentation; 4) The learner preparation of pre-project stage; 5) Writing the schemes and skills development training; 6) The introduction of presentation skill and sharing ideas through social network; 7) The presentation skill training in group and individual; 8) Multi-channel evaluation; 9) Roles and responsibilities during project stage and communication; and 10) Instructor evaluation, feedback, congratulations to learners who achieve successful project.

The researchers adopted the framework from Department of Curriculum and Instruction Development, Ministry of Education [4] to divide Project-Based Learning into 3 steps and rearrange the factors as follows:

A. Pre-Project Stage

This stage contained the following components:

- The 3rd component: The introduction of presentation skill and sharing ideas through social network
- The 5th component: The learner preparation of pre-project stage
- The 9th component: Writing the schemes and skills development training
- The 10th component: The presentation skill training in group and individual

B. While / During-Project Stage

This stage consisted of the following components:

- The 2nd component: Roles and responsibilities during project stage and communication
- The 6th component: Readiness of performing duties of individual or group of people which communicate through social network
- The 8th component: Guidance and agreement which learner submit voluntarily

C. Post-Project Stage

This stage consisted of the following components:

- The 1st component: Evaluations, document management system, presentation
- The 4th component: Instructor evaluation, feedback, congratulations to learners who achieve successful project
- The 7th component: Multi-channel evaluation

VII. RESEARCH RESULT SUMMARY

A. Evaluation of Each Item as Regards the Components of Project-Based Learning on Social Network with a Case Study on Presentation Skill Course

According to the study of 45 items individually which affect Project-Based Learning on social network with a case study on Presentation Skill course, it was found that the mean score was between 3.73 and 4.14 with standard deviation of between .680 and .909. This means that the all items as regards the components of Project-Based Learning on social network with a case study on Presentation Skill course were at high level and the deviation in each item was quite low.

B. Analysis of the Components of Project-Based Learning on Social Network with a Case Study on Presentation Skill Course

There were 10 components which are important for Project-Based Learning on social network with a case study on

Presentation Skill course. The factor meaning of each component could be given below.

- The 1st component: Evaluations, document management system, presentation
- The 2nd component: Roles and responsibilities during project stage and communication
- The 3rd component: The introduction of presentation skill and sharing ideas through social network
- The 4th component: Instructor evaluation, feedback, congratulations to learners who achieve successful project
- The 5th component: The learner preparation of pre-project stage
- The 6th component: Readiness of performing duties of individual or group of people which communicate through social network
- The 7th component: Multi-channel evaluation
- The 8th component: Guidance and agreement which learner submit voluntarily
- The 9th component: Writing the schemes and skills development training
- The 10th component: The presentation skill training in group and individual

C. Correlation Coefficients of the Components of Project-Based Learning on Social Network with a Case Study on Presentation Skill Course

The correlation coefficients between 10 components and 45 items which affect Project-Based Learning on social network with a case study on Presentation Skill course ranged from .424 and .935. The correlation coefficients among 10 components which affect Project-Based Learning on social network with a case study on Presentation Skill course ranged from .591 and .912. The correlation was high whereas the correlation coefficients for internal factors ranged from .002 and .070, with low correlation. The highest item which affects Project-Based Learning on social network with a case study on Presentation Skill course was shared communication channel through facebook. The lowest items were evaluation of skill development as well as presentation during project and self-evaluation by learners to reflect feelings and ideas through facebook.

D. Independent Variables or Predictors Which Are Selected into Regression Equation or Forecast Equation

There were 10 independent variables or predictors which are selected into regression equation or forecast equation. When the forecast equation is made for the analysis of components of Project-Based Learning on social network with a case study on Presentation Skill course, it is as follows:

$$Y = .780(\text{Factor 1}) + .603(\text{Factor 2}) + .686(\text{Factor 3}) + .591(\text{Factor 4}) + .741(\text{Factor 5}) + .912(\text{Factor 6}) + .632(\text{Factor 7}) + .824(\text{Factor 8}) + .689(\text{Factor 9}) + .686(\text{Factor 10})$$

when

- Y = Project-Based Learning on social network with a case study on Presentation Skill course
- Factor 1 = Evaluations, document management system, presentation
- Factor 2 = Roles and responsibilities during project stage and communication
- Factor 3 = The introduction of presentation skill and sharing ideas through social network
- Factor 4 = Instructor evaluation, feedback, congratulations to learners who achieve successful project
- Factor 5 = The learner preparation of pre-project stage
- Factor 6 = Readiness of performing duties of individual or group of people which communicate through social network
- Factor 7 = Multi-channel evaluation
- Factor 8 = Guidance and agreement which learner submit voluntarily
- Factor 9 = Writing the schemes and skills development training
- Factor 10 = The presentation skill training in group and individual

In summary, it was found that these 10 independent variables or predictors were correlated with Project-Based Learning on social network with a case study on Presentation Skill course. The coefficients ranged between .591 and .912 with high correlation. All 10 coefficients formed cumulative predictive power of 60.00% and standard error of estimate of 10.00.

VIII. RESEARCH DISCUSSIONS

A. Pre-Project Stage

The research results showed that the components of Project-Based Learning on social network with a case study on Presentation Skill course for the Pre-Project Stage consisted of the 3rd component: The introduction of presentation skill and sharing ideas through social network, the 5th component: The learner preparation of pre-project stage, the 9th component: Writing the schemes and skills development training and the 10th component:

The presentation skill training in group and individual; The results from this study are in compliance with Department of Curriculum and Instruction Development, Ministry of Education [4] in that Pre-Project Stage must equip learners with skills so that learners can develop and improve their skills. The results are also in accordance with the research by Brien [5] who found that planning and skill training can help learners achieve their learning objectives according to their desired direction, especially for those who are new to Project-Based Learning or those with low learning achievement. However, it was found that in the Pre-Project Stage learners must be given advice and guidance from instructors so that they can share ideas, brainstorm and present their ideas on social network.

That means learners need to develop their skill and this requires training.

B. While / During-Project Stage

The research results showed that the components of Project-Based Learning on social network with a case study on Presentation Skill course for the While / During-Project Stage consisted of the 2nd component: Roles and responsibilities during project stage and communication, the 6th component: Readiness of performing duties of individual or group of people which communicate through social network, and the 8th component: Guidance and agreement which learner submit voluntarily. These research results are in accordance with Department of Curriculum and Instruction Development, Ministry of Education [4] in that in While / During-Project Stage learners would apply their skills and language in the activities in a step-by-step manner. Learners will become active in doing their own things and instructors will become facilitators instead.

C. Post-Project Stage

The research results showed that the components of Project-Based Learning on social network with a case study on Presentation Skill course for the Post-Project Stage consisted of the 1st component: Evaluations, document management system, presentation, the 4th component: Instructor evaluation, feedback, congratulations to learners who achieve successful project, and the 7th component: Multi-channel evaluation. These research results are in accordance with Department of Curriculum and Instruction Development, Ministry of Education [4] in that in Post-Project Stage learners will assess themselves, their peers and their instructors. There will be congratulations to those who succeed in their projects and the documents will be written up and stored in a systematic manner.

IX. SUGGESTIONS

A. Suggestions from the Research Results

According to the research results, the 1st component of Project-Based Learning on social network is Evaluations, document management system, presentation. Therefore, instructors should have more ways to assess students such as self-assessment form, peer assessment form for classroom interaction. Instructors could assess the students during or after the project activities. Moreover, project participants could also have their form to assess the project organizers.

It was found that the 2nd component of Project-Based Learning on social network is Roles and responsibilities during project stage and communication. Therefore, instructors and learners should have their clear roles and responsibilities during the project management so that they can have a clear communication and to understand each other about their responsibilities. Researchers suggest using social network websites (such as facebook) to communicate to everybody and post comments.

The research results showed that the 3rd component of Project-Based Learning on social network is The introduction of presentation skill and sharing ideas through social network. Therefore, instructors should guide their students on how to

give presentation on social network. There can be samples of demonstration. Afterwards, the instructors and the learners could share their ideas about the skills which the instructors taught. This could be done through social network.

B. Suggestions for Further Research

1. There can be research on the development of Project-Based Learning on social network with a case study on Verbal Language for Communication.
2. There can be a comparative study between Project-Based Learning on social network and Project-Based Learning in classroom.
3. There can be an analytical study of components of Project-Based Learning on social network which are applicable in Presentation Skill course.

REFERENCES

- [1] Quick PC Extreme, "Social networking: How it reforms education", [Online], Available: <http://www.quickpcextreme.com/blog/?p=4985>, 2010, [Retrieved 20 January, 2011].
- [2] MGT News, Faculty of Management Sciences, Songkhla Rajabhat University, "Online social network and its advantage for collaborative learning", [Online], Available: <http://thanapat.blogspot.com/2009/07/social-network.html>, [Retrieved 25 February, 2011].
- [3] Registration and Evaluation Division, King Mongkut's University of Technology Thonburi, "Information about Curriculum and Description fo ETM 205 Presentation Skill course", [Online], Available: <http://regis.kmutt.ac.th/main.php>, [Retrieved 17 December, 2011]
- [4] Department of Curriculum and Instruction Development, Ministry of Education, Ways to Assess and Measure Learning Achievement according to Basic Education Curriculum 2544 B.E. Bangkok: Ministry of Education, 2011.
- [5] D. P. Brien, "The Teaching and Learning Processes Involved in Primary School Children's Research Projects.", Doctoral dissertation, University of New South Wales, 1995.

APPENDIX

Screenshots of social network website (facebook) which were used for Project-Based Learning in the academic year 2553 B.E. (2010)



Figure 1. Screenshot of main page for Project-Based Learning (2010)

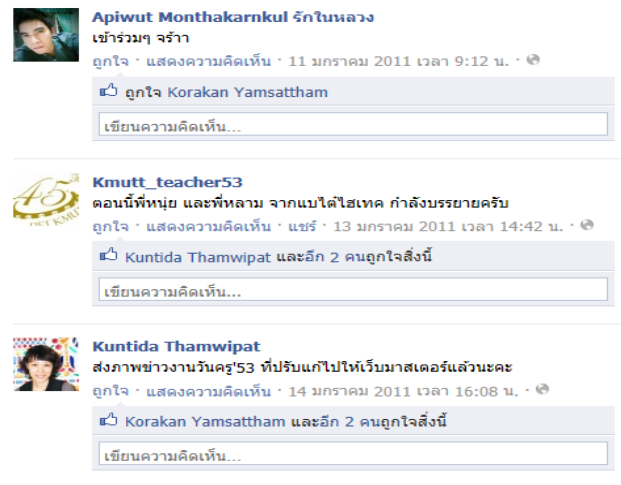


Figure 2. Screenshot of comments between learners and instructors

Screenshots of social network website (facebook) which were used for Project-Based Learning in the academic year 2554 B.E. (2011)



Figure 3. Screenshot of main page for Project-Based Learning (2011)



Figure 4. Screenshot of comments between learners and instructors

AUTHORS PROFILE

Associate Professor Dr. Kuntida Thamwipat is Associate Dean in Information and Student Affairs, Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi

Napassawan Yookong is a graduate student of Department of Educational Communications and Technology, Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi, Thailand

A Database Creation for Storing Electronic Documents and Research of the Staff

A Case of Department of Educational Communications and Technology

Pornpapatson Princhankol

Department of Educational Communications and
Technology,
King Mongkut's University of Technology Thonburi,
Bangkok, Thailand

Siriwan Phacharakreangchai

Department of Educational Communications and
Technology,
King Mongkut's University of Technology Thonburi,
Bangkok, Thailand

Abstract—The research study aims at creating the database for storing Electronic Documents and Research of the staff in the Department of Educational Communications and Technology, evaluating its quality and measuring the satisfaction of the database by users. The sample subjects are 14 instructors and officers in the Department of Educational Communications and Technology, King Mongkut's University of Technology Thonburi, in the second term of academic year 2012. The research tool is the database for storing Electronic Documents and Research of the staff in Department of Educational Communications and Technology. According to the evaluation by 3 experts in the area of Education Quality Assurance, the values of mean score and standard deviation are 4.51 and 0.38, respectively. This implies that the corresponding assurance quality of the database is very good. The media quality evaluation by 3 media specialists shows that its mean score is 4.58 and its standard deviation is 0.46. The evaluation result consequently indicates a very good quality of media. In terms of satisfaction on the database, the results obtained using questionnaires shows high satisfaction scores. The mean score is 4.15 and the standard deviation is 0.55 and this can be concluded that the database can be used effectively in the department.

Keywords—database creation; electronic documents; research.

I. BACKGROUND AND RATIONALE

Research is an important duty of higher education institutions and universities along with other duties such as teaching, producing graduates and distributing knowledge to the society. Research is important because it builds up knowledge based for instruction and distribution of knowledge, especially in this current society which gains a lot of impact from globalization and advances in technology. Knowledge has played a major role and it could be said that we are now in the era of knowledge management, in other words, instead of managing information, we need to know how to manage various bodies of knowledge. Research is one approach to such knowledge building. Therefore, research is the basis for knowledge management. In higher education institutions, there are many kinds of activities related to knowledge in terms of exploration, maintenance, distribution and application. Research is not only one of the main responsibilities but also has a role in creating and promoting other relevant missions of the universities such as instruction, graduate production and

academic scholarship. Therefore, research is necessary for the economic development and the national innovation. As a result, higher education institutions tend to focus more on research, in terms of both quality and mechanism improvement for research quantity [1].

In 2009, King Mongkut's University of Technology Thonburi was approved by the Office of the Higher Education Commission to be 1 out of 9 Research Universities of Thailand. Therefore, KMUTT continues to do research in order to meet the standards and to help increase the rankings of Thai universities to be higher in the Times Higher Education QS World University Rankings. Besides research of high quality, the production of high quality graduates is another main responsibility of universities with the aim to create graduates who are both bright and virtuous. KMUTT was funded to improve its potentials in research, innovation and identity [2].

King Mongkut's University of Technology Thonburi always plans to do research and improve the academic excellence through the building up and the application of knowledge in many dimensions. Besides, KMUTT aims at adjusting and renovating the organizations continuously so that the distribution of academic knowledge could be done and integrated with technology and social sciences which are taught on campus.

Faculty of Industrial Education and Technology has a policy to focus on the academic excellence in vocational education, industrial technology and related areas. There are many levels of professional development ranging from undergraduates, Master's and doctoral students. There are also research, training and workshops for the community and this reflects the leadership in vocational education and industrial technology. Moreover, there are links among universities, vocational institutions, organizations, foundations, and professional bodies related to this discipline both from national and international scales.

The policy of the research could be seen as 3 ways: (a) to promote and support research of the staff, students and the personnel in all levels and all disciplines in the Faculty of Industrial Education and Technology; (b) to supervise and oversee the quality of standards, visions, strategies, and mechanism to improve the research works in an efficient

manner; (c) to promote and distribute the research findings so that there can be a systematic application to reflect the university policy.

Faculty of Industrial Education and Technology always realizes the importance of research and the quality of research output to link or integrate various dimensions of academic excellence, which are research, instruction, technology and professional development. This could lead to the strength in postgraduate educational level. To increase the standards of vocational teachers and technologists is one mission of KMUTT. Hence, there are many research works on the building up of knowledge and innovation for the nation. However, there are no official researchers in the Faculty of Industrial Education and Technology because lecturers and postgraduate students have done research continuously and have published as well as presented their works in both journals and conferences in both Thailand and other countries. There are 3 kinds of research as in engineering research, learning research and general research [3].

In the past, the research works of the staff in the Department of Educational Communications and Technology were stored in the paper format and there were many difficulties. It was inconvenient for searching and retrieval and unresponsive from time to time. Therefore, the researchers would like to create a database for storing electronic documents and research of the staff in the department in order to improve the storage system of the staff.

II. RESEARCH OBJECTIVE

1. To create a database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology.
2. To evaluate the quality of the database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology.
3. To measure the satisfaction of the database users about the electronic documents and research of the staff in the Department of Educational Communications and Technology.

III. EXPECTED OUTCOMES

1. There would be a database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology.
2. The database for storing electronic documents and research of the staff could be used in the quality assurance process for the Department of Educational Communications and Technology.

IV. RESEARCH SCOPE

A. Population in This Study

The population in this study comprised 14 lecturers and administration staff members from the Department of Educational Communications and Technology, King

Mongkut's University of Technology Thonburi in the second term of the academic year 2012.

B. Research Scope

1. The database would be created for storing electronic documents and research of the staff in the Department of Educational Communications and Technology.
2. The electronic documents and research of the staff in the Department of Educational Communications and Technology would be collected and stored (from 2006 to 2011).
3. There would be links to these electronic documents and research of the staff in the Department of Educational Communications and Technology.

C. Variables

The independent variable in this study was the database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology.

The dependent variables in this study were the quality and the satisfaction of the users towards the database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology.

V. TOOL USED IN THIS RESEARCH

1. The database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology
2. The quality evaluation form for the database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology
3. The questionnaire on the satisfaction towards the database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology

The statistical approach to this research includes mean, standard deviation and index of congruence (IOC).

VI. RESEARCH RESULTS

Table I shows the overall quality evaluation by the experts in the area of Education Quality Assurance.

TABLE I. QUALITY EVALUATION

Items for Assessment	μ	σ	Quality Level
1 Input	4.53	0.35	Very Good
2 Process	4.53	0.35	Very Good
3 Output	4.47	0.46	Good
Mean score	4.51	0.38	Very Good

Table II shows the overall quality evaluation by the experts in the area of Media System.

TABLE II. QUALITY EVALUATION

Items for Assessment	μ	σ	Quality Level
1 Input	4.53	0.46	Very Good
2 Process	4.67	0.57	Very Good
3 Output	4.53	0.35	Very Good
Mean score	4.51	0.38	Very Good

Table III shows the overall satisfaction of the users towards the database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology.

TABLE III. USER SATISFACTION

Items for Assessment	μ	σ	Satisfaction Level
1 Input	4.59	0.55	The Highest
2 Process	4.46	0.54	High
3 Output	4.49	0.57	High
Mean score	4.51	0.55	The Highest

VII. RESEARCH RESULT SUMMARY

A. Result from the Creation of Database

According to the result from the creation of database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology, this research met the expectations as it was developed and revised according to the suggestions and advice given by the experts. The data were analyzed and this database was already installed in the Department of Educational Communications and Technology.

B. Result from the Quality Evaluation of the Database by the Experts in Education Quality Assurance

The results from the quality evaluation of the database by the experts in the area of Education Quality Assurance showed that the input was 4.53 on average with standard deviation of 0.35 or at very good level. As for the process, the mean score was 4.53 with standard deviation of 0.35, or at very good level. As for the output, the mean score was 4.47 with standard deviation of 0.46 or at good level. Therefore, the overall quality as assessed by the experts was 4.51 with standard deviation of 0.38. It means that the quality in terms of education quality assurance was at very good level.

C. Results from the Quality Evaluation of the Database by the Experts in Educational Media

The results from the quality evaluation of the database by the experts in educational media showed that the mean score for input was 4.53 with standard deviation of 0.46 or at very good level. As for the process, it was 4.67 on average with standard deviation of 0.57, or at very good level. In terms of output, the mean score was 4.53 with standard deviation of 0.35, or at very good level. Therefore, the overall mean score

by the experts in educational media was 4.58 with standard deviation of 0.46. This means that the quality in terms of educational media was at very good level.

D. Results from the Users' Satisfaction towards the Database

The results from the users' satisfaction towards the database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology showed that the mean score for input was 4.59 with standard deviation of 0.55, or at the highest level. As for the process, the mean score was 4.46 with standard deviation of 0.54, or at high level. In terms of output, the mean score was 4.49 with standard deviation of 0.57, or at high level. Therefore, the overall mean score for the users' satisfaction was 4.51 with standard deviation of 0.55. This means that the users expressed the highest satisfaction level.

VIII. RESEARCH DISCUSSIONS

A. Quality Evaluation of the Database

The results from the quality evaluation of the database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology showed that the quality in terms of quality assurance was at very good level and that the quality in terms of educational media was also at very good level. This means that the database was effective and suitable for storing electronic documents and research of the staff so that the documents could be used for quality assurance and for those who would like to find research works related to their work. Since the research documents could be dated back 5 years ago (2006-2010), a systematic approach will facilitate document storage and retrieval. This means that it could be searched in a short time and there will be more physical space for other documents. This view complies with the research work by Khonsan Reetanont [3] in that the existing educational quality assurance system in the Department of Educational Communications and Technology still depends on paper and it results in a large amount of paper. It takes longer time to find a document and it is inconvenient. Moreover, certain information might be lost or damaged and it could not be retrieved. The quality assurance results might not be satisfactory. Therefore, the new database system can provide the following features:

1. The electronic documents and research of the staff would be stored in a systematic order so that it is more convenient and quicker to retrieve the file.

2. There would be a decrease in physical space for storing the data. Moreover, it could prevent data loss or damaged documents.

3. The staff member could edit, modify and keep the data up to date to meet the demands of the users.

B. Users' Satisfaction towards the Database

As for the users' satisfaction towards the database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology, it was found that the users expressed the highest level for the input because it was convenient and quick to put the data into the database. It was also quick to edit, modify, check and verify

the data. Moreover, there are useful features such as automatic input and keywords so that the users can gain access to the database in a short time (as this point is the most satisfactory for users). To be able to gain access quickly means that the data could be kept up-to-date and that it could be used at anytime.

As for the process, the users expressed high level of satisfaction, especially towards the system procedure because it was easy to follow and there was no confusion. The system was table without error, resulting in less time for operation. There was also a system for feedback and comment from users so that the system could be modified to meet the demands of users. This view was similar to Jidapas Samphansomphot and Chaiyong Uprasitwong [4] in that the use of electronic document storage system could benefit to the uttermost and this would motivate the staff members to make the best use of the system. However, the compatibility must be taken into account for each department because the data could be changed all the time and each change in the system must improve the effectiveness of the data storage system.

As for the output, it was found that the users' satisfaction was at high level because the data display was quick, accurate and comprehensible. The data met the demands of the users and they were presented in an attractive manner with graphics, colors and illustrations. Moreover, the format and the text as well as the layout help embellish the display. According to the questionnaire results, it was found that the population expressed a high level of satisfaction because this database could reduce the operation time and the data were displayed in a comprehensible manner. The system was designed to store many kinds of information, including the data in the format of hard copy and digital file. It was also designed to reflect the actual structure of the filing system so that it is easy to gain access in the future. This operation was in compliance with the research by Orathai Thammasiri [5] in that the electronic document storage system is an extended version of the electronic catalogue system which involves the document operation. At the present time, there is an increase in the users, both centrally and locally. This could help solve many problems and it results in more convenience and data exchange.

IX. SUGGESTIONS

A. Suggestions for the Application of Research Results

According to the research results, it could be found that the creation of the database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology was suitable and could be a model for other departments or organizations to build a database for storing their own data. The essential information for those who are interested in building up a database could be as follows:

1. Data preparation must be done through a systematic classification. Then the documents could be converted into PDF format and the title must be named in accordance with the topic so that it is convenient to find the file when it is uploaded.

2. Database design must be based on the analysis of data storage preference. The table must be designed after an analysis of the way the users store their data so that it is convenient to

gain access to the file. The interface must be designed to interact with the users. The main focus should be on convenience for each component.

3. Virtual server installation could be done using AppServ software (Version 2.54a for Windows or higher) in order to simulate how the database would be used in the real situation. Since AppServ software contains compiler for PHP, MySQL and Server system, the operation could reflect the actual usage on the real server.

4. Centre for functionality could be made so that it is more convenient to use and edit commands when there is an error message. Moreover, it could reduce the delay when the system is used simultaneously.

According to the research results, the input, the process and the output are the main components for the database. Therefore, for those who would like to create a database, the procedure and the relationship among these three components must be taken into account for that the system could achieve its quality.

B. Suggestions for Further Research

1. There can be research on the creation of database for storing electronic documents and research of the Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi.

2. There can be a study on the creation of database regarding the training seminars of the lecturers and staff in the Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi.

REFERENCES

- [1] S. Khowilaikul, "Strategies to Develop Research Culture in Higher Education Institutions," *Journal of Education (Local)*, Chulalongkorn University, vol. 29, no. 1 (July-October 2000), 2000, pp. 8-18.
- [2] King Mongkut's University of Technology Thonburi, Annual Report 2009, Bangkok: Amarin Printing PLC, 2009, p. 5
- [3] K. Reetanont, "The Development of Information Technology System for Educational Quality Assurance: A Case Study of the Department of Computer and Educational Technology," a Master's thesis in Computer and Information Technology, Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi, 2006, p. 69.
- [4] J. Samphansomphot, and C. Uprasitwong, "Information System for Management", [Online], Available: <http://courseware.payap.ac.th/docu/sc312/lesson5/lesson05.htm> [Retrieved: 20 March, 2010].
- [5] O. Thammasiri, "Electronic Document Storage", [Online], Available: http://kmcenter.rid.go.th/kmc13/d_b/judgeb.htm [Retrieved: 14 February, 2011].

AUTHORS PROFILE

Associate Professor Dr. Pornpapatson Princhankol is Assistant Dean in Educational Quality Assurance, Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi

Siriwan Phachrakreangchai is a graduate student of Department of Educational Communications and Technology, Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi, Thailand

APPENDIX

Screenshots of the database for storing electronic documents and research of the staff in the Department of Educational Communications and Technology:

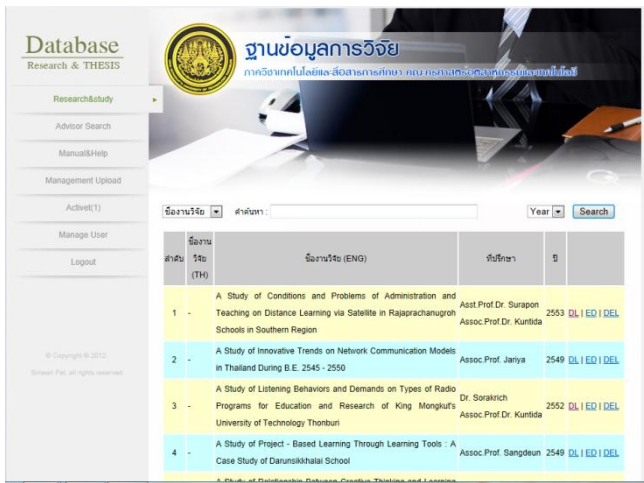


Figure 1. Screenshot of database main page



Figure 3. Screenshot of database access through manual and help

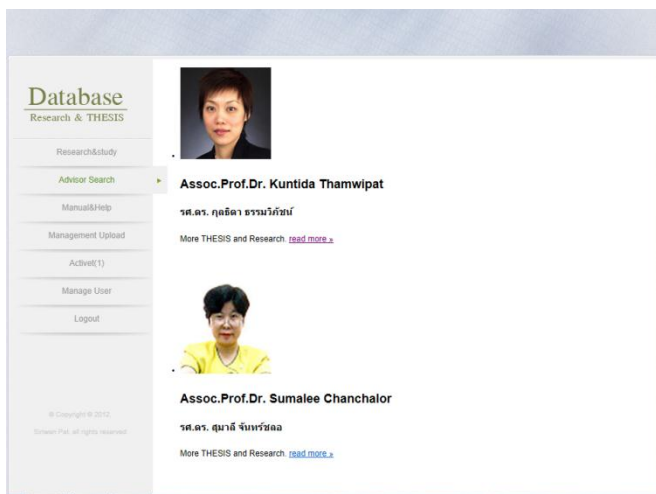


Figure 2. Screenshot of database access through advisor search

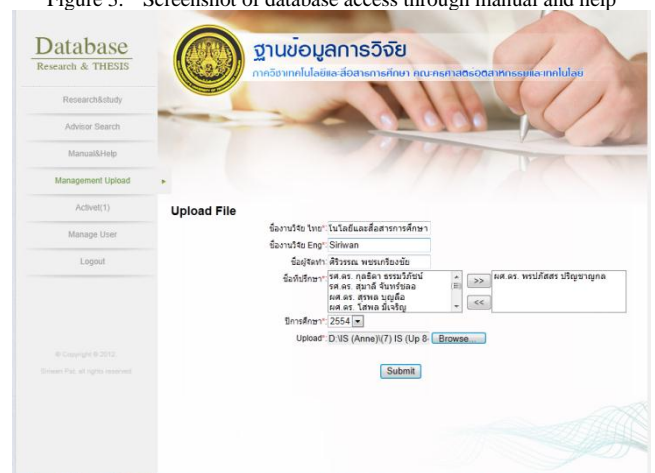


Figure 4. Screenshot of upload page to store documents in the database