IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 5 Issue 4

April 2014

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

SAI

www.ijacsa.thesai.org

IJACSA

WHERE WISDOM SHARES

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

OAIster getCITED Google Scholar BETA BASE Bielefeld Academic Search Engine ULRICHSWEB GLOBAL SERIALS DIRECTORY arXiv.org

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS IET InspecDirect INDEX COPERNICUS INTERNATIONAL WorldCat Window to the world's libraries Microsoft Academic Search EBSCO HOST Research Databases

Editorial Preface

From the Desk of Managing Editor…

It is our pleasure to present to you the April 2014 Issue of International Journal of Advanced Computer Science and Applications.

Today, it is incredible to consider that in 1969 men landed on the moon using a computer with a 32-kilobyte memory that was only programmable by the use of punch cards. In 1973, Astronaut Alan Shepherd participated in the first computer "hack" while orbiting the moon in his landing vehicle, as two programmers back on Earth attempted to "hack" into the duplicate computer, to find a way for Shepherd to convince his computer that a catastrophe requiring a mission abort was not happening; the successful hack took 45 minutes to accomplish, and Shepherd went on to hit his golf ball on the moon. Today, the average computer sitting on the desk of a suburban home office has more computing power than the entire U.S. space program that put humans on another world!!

Computer science has affected the human condition in many radical ways. Throughout its history, its developers have striven to make calculation and computation easier, as well as to offer new means by which the other sciences can be advanced. Modern massively-paralleled super-computers help scientists with previously unfeasible problems such as fluid dynamics, complex function convergence, finite element analysis and real-time weather dynamics.

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.

Lastly, we would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations.

We hope that materials contained in this volume will satisfy your expectations and entice you to submit your own contributions in upcoming issues of IJACSA

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 5 Issue 4 April 2014
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

CONTENTS

A Proposed Architectural Model for an Automatic Adaptive E-Learning System Based on Users Learning Style

Adeniran Adetunji
Physics Department,
The Polytechnic, Ibadan.
Ibadan, Nigeria.

Akande Ademola
Physics Department,
The Polytechnic, Ibadan.
Ibadan, Nigeria.

***Abstract*—It has been established through literature that, if an e-learning system could adapt to learning characteristics of learners, it will increase learning performance and content knowledge acquisition of learners. This paper is a *basic research* work for knowledge that lay down a foundation for application and implementation. We reviewed trends in adaptive e-learning system development, make an expository on learning-style models towards learners' learning character and propose an Architectural model of Automatic Adaptive E-learning System (AAeLS) based on learning-style concept/models. The concept it to model an e-learning system that will automatically adapt to learning preference of users', the system learn about users' learning style while the user learn the material content of the system; thus the learning process in two ways, the system is learning when the user is learning. We recommend further work on implementation and testing of the model, in an *applied research*.**

Keywords*—*E-Learning; Learning Style; Adaptation; AAeL

I. INTRODUCTION

Any educational platform is aimed at providing students with required information to increase their active knowledge about a particular subject. However, learning process is a variable that depends on the prior knowledge, motivation and needs of individual learners, [7]. This understanding poses a problem that emphasizes the importance of developing an adaptive system, which considers the individual needs of learners towards an effective learning process and acquisition of knowledge.

This paper review the concept of learning style and its variation in students and proposes an architectural model for an adaptive e-learning (AAeL) system based on differential concept of learning style. An adaptive learning system is one that is able to provide content information in a way that adapts to the prior knowledge and skill of learners, their learning capabilities, preference or style, their performance level and knowledge state, interests, personal circumstances and motivation, [7].

Developing training material and making them accessible on the internet is not enough, it is more important that the knowledge materials are tailored towards various learning characteristics of learners for example, their *learning styles.*

A recent trend in technology enhanced learning is integrating adaptive educational system into e-Learning. The rationale is that adapting courses to the learning preferences of the students has a positive effect on the learning process, leading to an increased efficiency, effectiveness and/or learner satisfaction, [5].

A common feature of an adaptive e-learning system is that they build a model of learner characteristics and use that model throughout the interaction with the learner, [2]. The aim is to provide the students the appropriate content at the right time, means that the system is able to determine the knowledge level, keep track of usage, and arrange content automatically for each student towards best learning result, [18]. Modritscher et al, 2004 idealize realization of adaptation in respect of learning and teaching process in an ideal e-learning system to depend on the following four elements:

Adaptive content aggregation: Depending on the learning and teaching style the system could offer different types of contend beginning with static information units to fully interactive elements like simulations, games or questionnaires. Besides, the content can be assembled with regard to different background domains, levels of detail or multimedia formats.

Adaptive presentation: The presentation of the content can be enhanced with additional, prerequisite, comparative explanations and all possible variants of these methods as well as sorting content units towards criteria like relevance to background knowledge, knowledge level, and the like. These techniques can be realised using technique like conditional text, stretch-text, page variants, fragment variants and frame-based methods.

Adaptive navigation: Navigation can be adapted in terms of global or local guidance and global or local orientation. Therefore, an e-learning environment could offer direct guidance as well as sorting, hiding and annotating links.

Adaptive collaboration support: The kind of technique, which can be offered by a network-based education system uses the system's knowledge about learners to form a collaborating group and offers or suggests communication within these learners using collaboration software.

II. Learning Style

"Learning styles are characteristic cognitive, affective and psychological behaviors that serve as relatively stable indicators of how learners perceive, interact with, and respond to the learning environment." [15].

When learning content is properly channeled and designed to match students' learning style, they learn best and learning process becomes more effective and efficient, [15.

Adapting learning based on learning style is the process of acquisition of knowledge that is peculiar to individual students, the attitude and behaviours of learners and determines the preferred modes of presentation of material contents in a most effective way.

Jorge Mota, in his work attributed the drawback in the general acceptability of e-Learning platform as an alternative form of training to traditional classroom as the inconsideration of the variability of learners learning style in the presentation of educational content in most of the e-learning material produced.

To design an e-learning system that integrates learning style of learners to provide adaptation in the system one needs to identify the differences in learners' attitude toward learning and addresses those differences at individual and group levels of learners.

Kolbs' Developmental theory of learning

Kolb in his postulates defined four dimensions of learning mode, conceptualize and identify learning abilities as follows:

(a) Concrete Experience (CE) (feeling), (b) Reflective Observation (RO) (reflection, watching), (c) Abstract Conceptualization (AC) (abstractness, thinking), and (d) Active Experimentation (AE) (action, doing). Learners, according to the model, must resolve a dialectical tension between immediate concrete experience and analytical detachment. In Kolb's model there are two learning continuums. Learners must choose a location between AC to CE on one continuum and AE to RO on the others. The combination of choices one makes between abilities indicates both a preference for ones ability over another and a preference for a specific construct or combination of abilities, namely, a learning style [10][11].

However, adaptive e-Learning system is a prospective platform to test and validate the effectiveness of *learning style* concept to learning outcome.

Felder-Silverman learning style model

With several learning style models in literature, Felder-Silverman learning style seems to be a preferred model as it classifies learners in broader groups and have a more detailed description about learners' leaning styles, [16].

Another main issue is that FSLSM is based on tendencies, saying that learners with a high preference for certain behaviour can also act sometimes differently, [16], this statement is a justification for the proposed architectural model of this paper. The Felder Model is most appropriate for hypermedia course ware, often used in learning style and advanced learning technology related research, [16].

TABLE I. Identifying Learning Styles From Pattern of Behaviour Using The Four Dimensions of FSLSM Description of Learning Styles

Active	Reflective
- Discuses, explain or test learned material. - In discussion forum, post more often in other to ask, discuss, and explain something. - Perform more self-assessment tests and more exercises as well as spend overall more time on exercise. - Spend very little time of studying examples since they prefer doing something by themselves rather than looking at how someone else has solved a problem.	- Thing about and work alone on learn material. - Participate passively in discussion forum and frequently reading the posting but only rarely posting by themselves. - They visit and spend more time on reading material like content objects as well as stay longer at outlines. - They tend to take longer on self-assessment tests as well as on the result page of self-assessments and exercises for reflecting on their results. - Expected to answer the same question in a self-assessment test less often twice wrong.

Sensing	Intuitive
- Prefer facts and data in learned materials - prefer examples, spend more time on examples - like to solve problems based on standard procedures, learn existing approaches and a high number of conducted self-assessment tests and exercises in order to check the acquired knowledge. - Patient with details, work carefully but slowly	- like to study abstract theories and their underlying meaning - learn from content objects and use examples only as supplementary material. Spend higher time on content objects and lower time on examples. - creative and like challenges - answer questions about developing new solutions, which require the understanding of underlying theories and concepts.

Visual	Verbal
- learn best from what they can see such as graphics images, and flow charts.	- Prefer to learn from words, regardless whether they are spoken or written. - Tend to like communicating and discussing with others. - high number of visits and

	postings as well as high amount of time spent in a discussion forum can indicate a verbal learning style. - Expected to visit reading material such as content objects more often.

Sequential	Global
- more comfortable with details - tend to go through the course step by step in a linear way.	- like to see the "big picture" and connections to other fields. - the outline of the course and the chapters are of interest global learners. - a high number of visits and more time spent on such chapter outlines as well as on the course overview page indicate a global learning style. - they are interested in relating and connecting topics to each other, this help them to interpret predefined solutions and develop new solutions. - tend to learn in large leaps, sometimes skipping learning objects and jumping to more complex material.

We are considering automation in determining students learning style in an adaptive e-learning system design. Student modeling can be in two different ways as distinguished by Brusilovsky, 1996. It can be either *Collaborative* or *automatic*. In the collaborative approach, learners provide explicit feedback which can be used to build and update a student model, such as filling out a learning style questionnaire. The automatic approach in the other hand builds and updates the student model automatically based on the users' behaviour and actions while using the system for learning, [16].

The problems of inaccurate self-conceptions of students are eliminated in the automatic approach. Moreover, it allows students to focus only on learning rather than additionally provide explicit feedback about their preferences, [16]. Another advantage of this approach is that, it analyses data from a specific time span rather than data which are gathered at one specific point of time. Common features of Learning Management Systems (LMSs) like content objects, outlines, examples, self-assessment tests, exercises and discussion forums a basis used by Sabine et al, 2009 in his automatic student modeling.

III. Adaptive Learning System Based On Learning Styles

Essaid et al, establishes that implementation of personalized e-learning system is a highly explored area of research in distance Web-based education. There have been identified differences in styles of learning and thus imperative to embed in out e-learning system, the ability to fit the learning process to these different needs of learners.

It has been generally accepted that the manner in which an individual choose or inclined to approach a learning circumstance is an influential factor to his/her performance and achievement of learning outcomes, [17]. The knowledge of the fact that there are various ways of approaching teaching and learning makes the difference in the design of an e-learning system with adaptive properties.

Yasir & Sami, 2011, proposes an approach to improve learning process through an adaptive hypermedia that provides adaptation in course content presentation based on students learning style. Experiment and statistical analysis indicates a better academic achievement in students taught with an e-learning system that adapts to their learning styles, [20].

Yasir & Sami, 2011, in their proposal for an Adaptive E-Learning Hypermedia System based on Learning Styles (AEHS-LS), presents the system based on three basic model: The ***domain model*** structured the subject matter, the knowledge of which to be learnt, the ***student model*** presents the initial knowledge level of the learner and ***the adaptation model*** provides the methods and techniques to select and present contents that fits learners knowledge state and learning preference.

Yang et al, et al 2013 proposed a personalized presentation module for developing adaptive learning system which is based on the field dependent/independent cognitive style and the eight dimensions of Felder-Silverman's learning style of students.

In most proposed adaptive learning system, only one or two dimensions of a learning style model are considered while developing the system, [21]. Yang et al, et al 2013, improved on this by developing their system based on both learning styles and cognitive styles to adapt the user interface and learning content for individual students and they took into consideration the full dimensions of a learning style model. They established through their experimental results that the proposed system could improve learning achievements of students, decreases their mental load and improve learning gains.

Farman et al, 2010 presents a concept of identifying and integrating learning styles and affective state of learners into e-learning system, aim to provide the adaptation need of individual learners. Most of the approaches to adaptive e-learning implementation based on learning style lacks dynamism in the model, [14].

Natalia et al, in their *Adaptive Hypermedia Architecture System* specifies instructional strategies and strategies for monitoring a learner's preference. They concluded through their experimental research work, that psychological and/or

pedagogical knowledge is required in the process of creating adaptive behaviour itself and recommended that the authors with experience in pedagogical psychology is allowed to design different types of strategies and apply these strategies to the applications. They also recommend that authors of application or psychologist are allowed to do the structuring of the application and organization of material content to correctly suit different learning styles, [14].

Meryem & Buket**,** 2002 concludes in their research paper that learning style do not have effect on the achievement of students in different learning environment. In my own interpretation, Meryem & Buket**,** 2002 is not declining the fact that "learning style consideration in learning process improves learning achievement" but claim that a particular learning style will pose the same effect in either e-learning platform or in a traditional classroom.

IV. AUTOMATIC ADDAPTIVE E-LEARNING SYSTEM (AAELS) ARCHITECTURE

Our proposed architecture as shown in figure 2 below;

The domain model: contains the material content of course to be learnt and is to be designed by both the teachers that handles ways of presenting/teaching the material content example;

- Presentation of content objects and outlines
- Explanations on objects and testing learned materials
- Presentation of facts and data on learning objects
- Presentation of theories and abstract modelling of learning objects
- Graphics, Image, and flow chart presentation of learning objects
- Textual presentation of learning objects
- Online discussion and chat forum on learned materials
- Solved examples and do-it-yourself exercises

e.t.c. .. .. ..

and author with psychological and pedagogical knowledge who categorizes the presentations into different learning styles. In this domain model we'll define different presentations of the same material content to meet needs of learners with various identified learning styles.

Leaners' model: the learner model of our proposed system, define three sub models; borrowing from**,** [20]. The three sub-models are:

The profile, the knowledge level and learning style of learners. The profile contains the static properties of learner, like the username, the password, unique ID, age, e-mail, the knowledge level and the learning style of learner will have dynamic properties in our model. We propose a feedback support from responses and activities at the presentation by the adaptation model which helps in modifying the user model.

Adaptation model: this defines the mode in which the learner's knowledge and learning style (learners' model) modify and determine content (domain model) presentation. The adaptation model is seamlessly updated and re-modified, as the feedback mechanism updates the users' knowledge level and learning preference. This is the automation concept of our model.

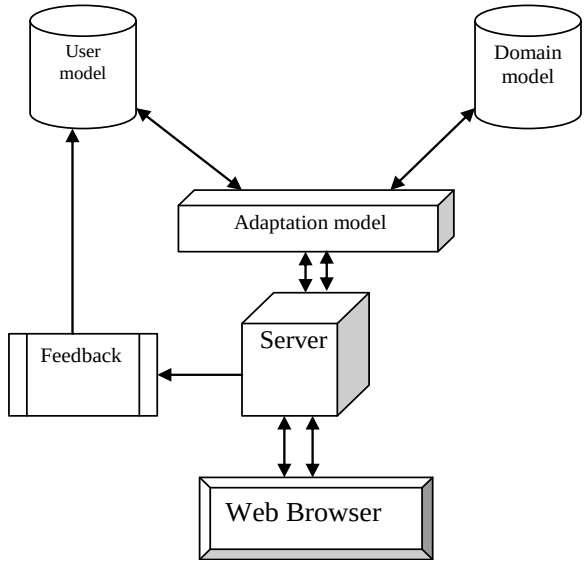


Fig. 1. AAEL-S Architecture

The proposed system can be implemented towards achieving the aim of an automatic adaptive e-learning system. The ***domain model*** structures the knowledge about the subject matter; the ***learners' model*** provides the description of the current state of learners' with respect to the knowledge about the domain to be learned, and the ***adaptation model*** implements the specified adaptation rules. The feedback mechanism in the proposed architecture enables the system to sense any change in the knowledge state and learning preference of learners and adaptation model is adjusted to fit the new needs of such learners, this process is going to be automatic without learners notice.

The system is to learn about the knowledge level and learning preference changes of learner while the learner learns the domain content from the system, the learning model is thus in two direction as depicted in figure 2 below:

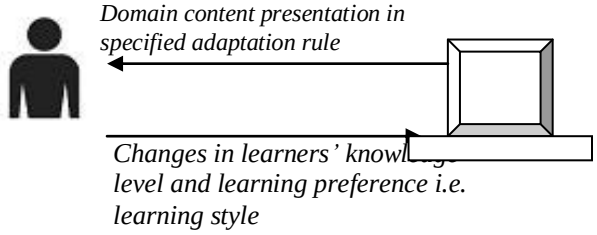


Fig. 2. two-way learning directions between learners and the AAEL-S.

V. CONCLUSION

This work is an expository on Adaptive E-Learning system. The system architecture proposed is a modified and improved model of a related work, [20]. The proposed system is recommended for further work on implementation and testing.

The aim is to provide basic information towards designing an e-learning system that adapts to learners' preference of "learning Style" *automatically* i.e. Automatic Adaptive E-Learning System (AAeLS). The adaptation by the modeled system does not require the user/learner to perform any preliminary activity before it gets information about their adaptive needs; the system does that automatically as user/learner does his/her own study through the e-Learning platform.

REFERENCES

[1] Boticario, J.G., Santos, O.C., van Rosmale P. (2005); "Issues in Developing Standard-based Adaptive Learning Management Systems." EADTU 2005 Working Conference: Towards Lisbon 2010: Collaboration for Innovative Content in Lifelong Open and Flexible Learning.

[2] Brusilovsky, P., Peylo, C.(2003); "Adaptive and Intelligent Web-based Educational Systems." International Journal of Artificial Intelligence in Education, 13 (2-4). Pp. 159-172.

[3] Brusilovsky, P.(1999); "Methods and Techniques of adaptive hypermedia, User Modeling and User-Adapted Interaction", 6, 1996, pp. 87-129.

[4] Dessislava Vassileva (2012); "Adaptive E-learning Content Design and Delivery Based on Learning Styles and Knowledge*", Serdica J. Computing 6 (2012), 207-252, Serdica Journal of Computing. Bulgarian Academy of Science Institute of Mathematics and Informatics.

[5] Elvira Popescu, Costin Badica and Lucian Moraret (2010); "Accommodating Learning Styles in an Adaptive Educational System", University of Craiova, A.I.Cuza 13, 200585 Craiova, Romania Email:popescu_elvira@software.ucv.ro, badica_costin@software.ucv.ro. Informatica 34(2010) 451-462.

[6] Essaid El Bachari, El Hassan Abdelwahed, Mohamed El Adnani; "Design of An Adaptive E-Learning Model Based ON Learner's Personality", Ubiquitous Computing and Communication Journal. Computer Systems Engineering Laboratory (LISI), Department of Engineering Science, Faculty of Science Semlalia, Cadi Ayyad University B.P. 2390, Bd My Abdellah, 40000, Marrakesh. {elbachari, abdelwahed, md-eladnani}@ucam.ac.ma.

[7] F. Kareal* and J. Klema (2006); "Adaptivity in e-learning", Department of Cybernetic, Faculty of Electrotechnics, Czeck Technical University, Technicka 2, 166 27 Prague 6, Czeck Republic.

[8] Farman Ali Khan, Sabine Graf, Edgar R. Weippl, A Min Tjoa (2010); "Implementation of Affective State and Learning Styles Tactics in Web-based Learning Management Systems", 2010 10th IEEE International Conference on Advanced Learning Technologies.

[9] Jorge Mota; "Using Learning Syteles and Neural Networks as an Approach to eLearning Content and Layout Adaptation", PRODE-Facult of Engineering of University of Porto, Portugal. Museu8bits@gmail.com.

[10] Kolb, D. (1976); "Learning style inventory." Boston: McBer and Company.

[11] Kolb, D. (1984); "Experiential learning." Englewood Cliffs, NJ: Prentice Hall.

[12] Meryem Yilmaz-Soylu & Buket Akkoyunlu (2002); "The Effect of Learning Sytles on Achievement in Different Learning Environments", The Turkish Online Journal of Education Technology 2002.

[13] Modritscher, F., Gutl, C., Garcia B., & Maurer, H. (2004). "Enhancement of SCORM to support adaptive e-learning within the scope of the research project AdeLE. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.6119&rep1&type=pdf.

[14] Natalia Stash, Alexandra Cristea, Paul De Bra; Faculty of Mathematics and Computer Science, Eindhoven University of Technology, The Nertherlands, {natash, acristea, debra}@win.tue.nl.

[15] Olga Mironova[1], Irina Amitan[2], Tiia Ruutmann[3] (2013); "Computer Science E-Course for Students with Different Learning Styles", [1&2]Informatics, Chair of Software Engineering, Tallinn University of Technology, Akadeemia tee St. 15A, Tallinn 12618, Estonia. Email: {olga.mironova, irina.amitan, juri.vilipold, merike.saar}@ttu.ee. [3]Faculty of Social Sciences, Department of Industrial Psychology, Estonian Centre for Engineering Pedagogy Tallin University of Technology, Akadeemia tee St. 3, Tallin 12618, Estonia Email: tiia.ruutmann@ttu.ee.

[16] [1]Sabine Graf, [3]Kinshuk, [4]Tzu-Chien Liu; "Identifying Learning Styles in Learning Management Systems by Using Indications from Students' Behaviour", [1]Women's Postgraduate College of Internet Technologies Vienna University of Technology, Veinna, Austria. graf@wit.tuwien.ac.at. [3]Shool of Computing and Information Systems Athabasca University, Athabasca, Canada. kinshuk@ieee.org. [4]Natioal Central University Graduate Institute of Learning and Instruction, Taiwan ltc@cc.ncu.edu.tw.

[17] Simon Cassidy* (2004); "Learning Styles: An overview of theories, models, and measures." Carfax Publishing, Educational Psychology Vol. 24, No. 4, August 2004.

[18] Vatcharaporn Esichaikul[1], Supaporn Lamnoi[2], Clemens Bechter[3]; "Student Modelling in Adaptive E-Learning Systems.", [1]School of Engineering and Technology Asian Institute of Technology, Thailand. E-mail: vatchara@ait.ac.th, [2]National Electronics and Computer Technology Center (NECTEC), Pathumthani 12120, Thailang. E-mail: suparportn.lamnoi@nectec.or.th, [3]Thammasat Business School, Thammasat University, Bangkok, Thailand. E-mail: bechter@gmail.com.

[19] Yasir Eltigani Ali Mustafa [1,2*] and Sami Mohamed Sharif[1](2011); "An approach to Adaptive E-Learning Hypermedia System based on Learning Styles (AEHS-LS): Implementation and Evaluation.", International Journal of Library and Information Science Vol. 3(1), pp. 15-28, January 2011. Available online http://www.academicjournals.org/ijlis, ISSN 2141-2537 ©2011 Academic Journals.

[20] Yang, T.-C., Hwang, G. –J., & Yang, S. J, -H. (2013); "Development of an adaptive learning system with multiple perspectives based on students' learning styles and congnitive styles. Educational Technology & Society, 16(4), 185-200.

An Optimized Analogy-Based Project Effort Estimation

Mohammad Azzeh
Faculty of Information Technology
Applied Science UniversityAmman,
Jordan POBOX 166

Yousef Elsheikh
Faculty of Information Technology
Applied Science University
Amman, Jordan

Marwan Alseid
Faculty of Information Technology
Applied Science University
Amman, Jordan

***Abstract*—Despite the predictive performance of Analogy-Based Estimation (ABE) in generating better effort estimates, there is no consensus on: (1) how to predetermine the appropriate number of analogies, (2) which adjustment technique produces better estimates. Yet, there is no prior works attempted to optimize both number of analogies and feature distance weights for each test project. Perhaps rather than using fixed number, it is better to optimize this value for each project individually and then adjust the retrieved analogies by optimizing and approximating complex relationships between features and reflects that approximation on the final estimate. The Artificial Bees Algorithm is utilized to find, for each test project, the appropriate number of closest projects and features distance weights that are used to adjust those analogies' efforts. The proposed technique has been applied and validated to 8 publically datasets from PROMISE repository. Results obtained show that: (1) the predictive performance of ABE has noticeably been improved; (2) the number of analogies was remarkably variable for each test project. While there are many techniques to adjust ABE, Using optimization algorithm provides two solutions in one technique and appeared useful for datasets with complex structure.**

Keywords—Cost Estimation; Effort Estimation by Analogy; Bees Optimization Algorithm

I. INTRODUCTION

Analogy-Based Estimation (ABE) has preserved popularity within software engineering research community because of its outstanding performance in prediction when different data types are used [1, 15]. The idea behind this method is rather simple such that the new project's effort can be estimated by reusing efforts about similar, already documented projects in a dataset, where in a first step one has to identify similar projects which contain the useful predictions [15]. The predictive performance of ABE relies significantly on the choice of two interrelated parameters: number of nearest analogies and adjustment strategy [8]. The goal of using adjustment in ABE is twofold: (1) minimizing the difference between a new project and its nearest analogies, and (2) producing more successful estimates in comparison to original ABE [2]. If the researchers read the literature on ABE, they will encounter large number of ABE models that use variety of adjustment strategies. Those strategies suffer from common problems such as they are not able to produces stable results when applied in different contexts as well as they use fixed number of analogies for the whole dataset [1]. Using fixed number of analogies has been proven to be unsuccessful in many situations because it depends heavily on expert opinion and requires extensive experimentation to identify the best *k* value, which might not be predictive for individual projects [2].

The aim of this work is therefore to propose a new method based on Artificial Bees Algorithm (BA) [14] to adjust ABE by optimizing the feature similarity coefficients that minimizes difference between new project and its nearest projects, and predicting the best *k* number of nearest analogies. The paper is structured as follows: Section 2 introduces an overview to ABE and adjustment methods. Section 3 presents the proposed adjustment method. Section 4 presents research methodology. Section 5 shows obtained results. Finally the paper ends with our conclusions.

II. RELATED WORKS

ABE method generates new prediction based on assumption that similar projects with respect to features description have similar efforts [8, 15]. Adjustment is a part of ABE that attempts to minimize the difference between new observation ($\hat{e}_i$) and each nearest similar observation (e_i), then reflects that difference on the derived solution in order to obtain better solution (e_t). Consequentially, all adjusted solutions are aggregated using simple statistical methods such as mean ($e_t = k^{-1}\sum_{i=1}^{k}\hat{e}_i$). In previous study [17] we investigated the performance of BA, on adjusting ABE and finding best *k* value for the whole dataset. This model showed some improvements on the accuracy, but on the other side it did not solve the problem of predicting the best *k* value for each individual project. In addition the solution space of BA was a challenge because there was only one common weight for all nearest analogies. The used optimization criterion (i.e. MMRE) was problematic because it was proven to be biased towards underestimation. For all these reason and since we need to compare our proposed model with validated and replicated models, we excluded this model from comparison later in this paper. This paper thereby attempts to solve abovementioned limitations.

In literature there is a significant number of adjustment methods that have been documented and replicated in previous studies. Therefore we selected and summarized only the most widely used strategies. Walkerden and Jeffery proposed Linear Size Adjustment (LSE) [16] based on the size extrapolation. Mendes et al. [12] proposed Multiple Linear Feature Extrapolation (MLFE) to include all related size features.

Jorgenson et al. [6] proposed Regression Towards the Mean (RTM) to adjust projects based on their productivity values. Chiu and Huang [4] proposed another adjustment based on Genetic Algorithm (GA) to optimize the coefficient αj for each feature distance based on minimizing performance measure. Recently, Li et al. [10] proposed the use of Neural Network (NN) to learn the difference between projects and reflects the difference on the final estimate. Further details about these methods and their functions can be found in [1].

Indeed, the most important questions to consider when to use such methods is how to predict the best number of nearest analogies (*k*). In recent years various approaches have been proposed to specify this number such as: 1) fixed number selection (i.e. *k*=1, 2, 3…etc) as in studies of [7, 11, 12, 16], 2) Dynamic selection based on clustering as in study of [2, 17]. 3) Similarity threshold based selection as in studies of [5, 9]. Generally, these studies except [2] use the same *k* value for all projects in the dataset which does not necessarily produce best performance for each individual project. On the other hand, the certain problem with [2] is that it does not include adjustment method but it predicts the best *k* value based on the structure of dataset.

III. The Proposed Method (OABE)

The proposed adjustment method starts with Bees Algorithm in order to find out, for each project: (1) the feature weights (*w*), and (2) the best *k* number of nearest analogies that minimize mean absolute error. The search space of BA can be seen as a set of *n* weight matrixes where the size of each matrix (i.e. solution) is $k \times m$. That means each possible solution contains weight matrix with dimension equivalent to the number of analogies (*k*) and number of features (*m*) as shown in Figure 1. The number of rows (i.e. *k*) and weight values are initially generated by random. Each row represents weights for one selected analogy and accordingly $\sum_{j=1}^{m} w_j = 1$. In each run the algorithm selects the top *k* nearest analogies based on the number of *k* weights in the search space. Then each selected analogy is adjusted with corresponding weights taken from the matrix *w* as shown Eq.1. The algorithm continues searching until the value of Mean Error (i.e. $MR = k^{-1}\sum_{j=1}^{k}\Delta_{ij}$) between new project and its *k* analogies is minimized. The optimized *k* value and weight matrix are then applied to Eqs. 1, 2 and 3 to generate new estimate. The new integration between ABE with BA will be called Optimized Analogy Based Estimation (hereafter OABE).

$$w = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ w_{k1} & w_{k2} & \cdots & w_{km} \end{bmatrix}$$

Fig. 1. Weight Matrix for one solution in the search space

$$\Delta_{ij} = \frac{1}{m}\sum_{j=1}^{m} w_{ij} \times (f_{tj} - f_{ij}) \quad (1)$$

$$\hat{e}_i = e_i + \Delta_{ij} \quad (2)$$

$$e_t = \frac{\sum_{i=1}^{k}(k+1-r_i)\times \hat{e}_i}{\sum_{i=1}^{k} i} \quad (3)$$

The setting parameters for AB have been found after performing sensitivity analysis on the employed datasets to see the appropriate values. Table I shows BA parameters, their abbreviations and initial values used in this study. Below we briefly describe the process of BA in finding best *k* values and the corresponding weights for each new project. The algorithm starts with an initial set of weight matrixes generated after randomly initializing *k* for each matrix. The solutions are assessed and sorted in ascending order after they are being evaluated based on *MR*. The best from 1 to *b* solutions are being selected for neighborhood search for better solutions, and form new patch. Similarly, a number of bees (*nsp*) are also recruited for each solution ranked from *b+1* to *u*, to search in the neighborhood. The best solution in each patch will replace the old best solution in that patch and the remaining bees will be replaced randomly with other solutions. The algorithm continues searching in the neighborhood of the selected sites, recruiting more bees to search near to the best sites which may have promising solutions. These steps are repeated until the criterion of stop (minimum *MR*) is met or the number of iteration has finished.

TABLE I. BA Parameters

Parameter	Description	Value
q	dimension of solution	(number of features +1)
n	represents size of initial solutions	100
u	number of sites selected out of *n* visited sites	20
b	number of best sites out of *s* selected sites	10
nep	number of bees recruited for best *b* sites	30
nsp	Number of bees recruited for the other selected sites	20
ngh	initial size of patches (*ngh*)	0.05

IV. Methodology

A. Datasets

The proposed OABE model has been validated over 8 software effort estimation datasets come from companies of different industrial sectors [3]. The datasets characteristics are provided in Table II which shows that the datasets are strongly positively skewed indicating many small projects and a limited number of outliers. It is important to note that all continuous features have been scaled and all observation with missing values are excluded.

TABLE II. DESCRIPTIVE STATISTICS OF THE DATASETS

Dataset	Feature	Size	Effort Data			
			Min	Max	Mean	Skew
Albrecht	7	24	1	105	22	2.2
Kemerer	7	15	23.2	1107.3	219.2	2.76
Nasa	3	18	5	138.3	49.47	0.57
Desharnais	12	77	546	23940	5046	2.0
COCOMO	17	63	6	11400	683	4.4
China	18	499	26	54620	3921	3.92
Maxwell	27	62	583	63694	8223.2	3.26
Telecom	3	18	23.54	1115.5	284.33	1.78

B. Performance measures

A key question to any estimation model is whether the predications are accurate, the difference between the actual effort (e_i) and the predicted effort ($\hat{e}_i$) should be as small as possible because large deviation will have opposite effect on the development progress of the new software project [13]. This section describes several performance measures used in this research as shown in Table III. Although some measures such as *MMRE*, *MMER* have been criticized as biased to under and over estimations, we insist to use them because they are widely used in commenting on the success of predictions [13].

TABLE III. ERROR MEASURES

Error Measure Name	Equation
Magnitude Relative Error	$MRE = \frac{\lvert e_i - \hat{e}_i \rvert}{e_i}$
Mean Magnitude Relative Error	$MMRE = N^{-1}\sum_i MRE_i$
Median Magnitude Relative Error	$MdMRE = median_i(MRE_i)$
Mean Magnitude of Error Relative to the estimate	$MMER = N^{-1}\sum_i \frac{\lvert e_i - \hat{e}_i \rvert}{\hat{e}_i}$
Mean Balanced Error (MBRE)	$MBER = N^{-1}\sum_i \frac{\lvert e_i - \hat{e}_i \rvert}{\min(e_i, \hat{e}_i)}$
Prediction Performance	$pred_l = \frac{100}{N} \times \sum_{i=1}^{N} \begin{cases} 1 & if\ MRE_i \le 0.25 \\ 0 & otherwise \end{cases}$

Interpreting these error measures without any statistical test can lead to conclusion instability, therefore we used *win-tie-loss* algorithm [8] to compare the performance of OABE to other estimation methods. We first check if two methods $method_i$; $method_j$ are statistically different according to the Wilcoxon test. If so, we update win_i; win_j and $loss_i$; $loss_j$after checking which one is better according to the performance measure at hand; otherwise we increase tie_i and tie_j. The performance measures used here are *MRE*, *MMRE*, *MdMRE*, *MMER*, *MBER* and $Pred_{25}$. Algorithm 1 illustrates the win-tie-loss algorithm [8]. Also, the Bonferroni-Dunn test is used to perform multiple comparisons for different models based on the absolute error to check whether there are differences in population rank means among more than populations.

Algorithm 1. Pseudocode of win-tie-loss algorithm between$method_i$ and $method_j$based on performance measure E [8]

1: $Win_i=0, tie_i=0, loss_i=0$
2: $Win_j=0, tie_j=0; loss_j=0$
3: **if** *Wilcoxon (MRE(*$method_i$*), MRE(*$method_j$*), 95) says they are the same* **then**
4: $tie_i = tie_i + 1;$
5: $tie_j = tie_j + 1;$
6:**else**
7: **if** *better(E(*$method_i$*), E(*$method_j$*))* **then**
8: $win_i = win_i + 1$
9: $loss_j = loss_j + 1$
10: **else**
11: $win_j = win_j + 1$
12: $loss_i = loss_i + 1$
13:**end if**
14: **end if**

V. RESULTS

This section presents performance figures of OABE against various adjustment techniques used in constructing ABE models. Since the selection of the best *k* setting in OABE is dynamic, there was no need to pre-set the best *k* value. In contrast, for other variants of adjustment techniques there was necessarily finding the best *k* value that almost fits each model, therefore we applied different *k* settings from 1 to 5 on each model as suggested by Li et al. [9] and the setting that produces best overall performance has been selected for comparison with other different models. Tables IV, V, VI, VII and VIII summarize the resulting performance figures for all investigated ABE models. The most successful method should have lower *MMRE*, *MdMRE*, *MMER*, *MBER* and higher $Pred_{25}$. The obtained results suggest that the OABE produced accurate predictions than other methods with quite good performance figures over most datasets.

TABLE IV. MMRE RESULTS

Dataset	OABE	LSE	MLFE	RTM	GA	NN
Albrecht	**40.2**	62.9	65.2	61.2	45.4	51.2
Kemerer	**39.6**	41.4	64.5	44.6	60.4	166.0
Desharnais	34.5	37.2	45.6	**33.4**	49.4	78.4
COCOMO	**50.1**	65.8	148.2	54.0	159.5	203.6
Maxwell	**41.7**	71.2	71.2	46.4	117.2	199.9
China	24.7	**20.9**	32.8	36.5	46.5	68.6
Telecom	**13.2**	15.4	36.7	15.2	39.1	78.9
Nasa	61.2	58.3	55.7	**54.9**	58.6	99.2

TABLE V. $PRED_{25}$ RESULTS

Dataset	OABE	LSE	MLFE	RTM	GA	NN
Albrecht	**44.6**	37.5	37.5	33.3	33.3	29.2
Kemerer	53.3	**60.0**	26.7	33.3	33.3	13.3
Desharnais	**48.2**	42.9	37.7	41.6	37.7	31.2
COCOMO	20.2	**31.7**	14.3	25.4	14.3	6.3
Maxwell	**34.4**	27.4	27.4	32.3	17.7	3.2
China	80.7	**82.4**	25.9	45.9	43.9	46.1
Telecom	**84.0**	77.8	55.6	77.8	61.1	22.2
Nasa	**50.0**	33.3	33.3	33.3	38.9	11.1

TABLE VI. MDMRE RESULTS

Dataset	OABE	LSE	MLFE	RTM	GA	NN
Albrecht	37.2	**29.7**	30.3	40.5	38.5	43.1
Kemerer	23.3	**21.3**	39.6	46.1	41.4	128.5
Desharnais	**26.3**	28.9	31.0	30.9	35.9	51.9
COCOMO	47.7	**38.0**	71.6	46.9	81.1	99.5
Maxwell	44.2	48.1	48.1	**41.0**	60.2	160.0
China	24.6	**22.6**	84.4	28.4	29.2	29.2
Telecom	**10.3**	13.4	20.0	12.6	18.7	58.4
Nasa	**25.8**	39.4	44.1	36.6	31.5	81.3

However, these findings are indicative of the superiority of BA in optimizing *k* analogies and adjusting the retrieved project efforts, and consequentially improve overall predictive performance of ABE. Also from the obtained results we can observe that there is evidence that using adjustment techniques can work better for datasets with discontinuities (e.g. Maxwell, Kemerer and COCOMO). Note that the result is exactly the "searching for the best *k* setting" result as might be predicted by the researchers mentioned in the related work. We speculate that prior Software Engineering researchers who failed to find best *k* setting, did not attempt to optimize this *k* value with adjustment technique itself for each individual project before building the model.

TABLE VII. MMER RESULTS

Dataset	OABE	LSE	MLFE	RTM	GA	NN
Albrecht	**38.6**	57.2	50.0	86.1	53.1	154.4
Kemerer	**51.3**	59.7	55.5	53.8	56.8	73.3
Desharnais	37.2	**35.2**	38.0	40.7	47.4	95.1
COCOMO	**58.0**	62.9	226.6	117.8	285.2	111.9
Maxwell	54.7	**48.3**	**48.3**	63.1	108.2	117.4
China	16.2	**14.8**	47.1	55.2	44.8	64.4
Telecom	**15.2**	18.2	27.1	16.1	26.5	357.9
Nasa	**44.4**	49.3	53.0	80.5	46.6	279.4

TABLE VIII. MBRE RESULTS

Dataset	OABE	LSE	MLFE	RTM	GA	NN
Albrecht	**61.2**	87.7	82.7	107.5	65.8	166.0
Kemerer	**57.5**	71.4	83.9	64.8	81.1	124.3
Desharnais	**40.4**	45.6	54.1	46.8	65.5	81.4
COCOMO	97.3	**92.9**	319.4	129.0	383.3	239.4
Maxwell	84.2	81.9	81.9	**74.3**	175.9	199.8
China	23.3	**23.0**	32.1	62.1	62.3	90.1
Telecom	**16.5**	16.9	39.7	17.4	42.6	73.0
Nasa	**71.1**	75.6	73.7	98.0	74.1	99.6

Furthermore, two results worth some attention while we are carrying this experiment: Firstly, the general trend of predictive accuracy improvements across all error measures, overall datasets is not clear this certainly depends on the structure of the dataset. Secondly, there is no consistent results regarding the suitability of OABE for small datasets with categorical features (as in Maxwell and Kemerer datasets) but we can insist that OABE is still comparable to LSE in terms of *MMRE* and $Pred_{25}$ and have potential to produce better estimates.

In contrast, OABE showed better performance than LSE for the other two small datasets (NASA and Telecom) that do not have categorical features. To summarize the results we run the win-tie-loss algorithm to show the overall performance. Figure 3 shows the sum of win, tie and loss values for all models, over all datasets. Every model in Figure 2 is compared to other five models, over six error measures and eight datasets. Notice in Figure 2 that except the low performing model on, the tie values are in 49-136 band. Therefore, they would not be so informative as to differentiate the methods, so we consult win and loss statistics to tell us which model performs better over all datasets using different error measures.

Apparently, there is significant difference between the best and worst models in terms of win and loss values (in the extreme case it is close to 119). The win-tie-loss results offer yet more evidence for the superiority of OABE over other adjustment techniques. Also the obtained win-tie-loss results confirmed that the predictions based on OABE model presented statistically significant but necessarily accurate estimations than other techniques.

Two aspects of these results are worth commenting: 1) The NN was the big loser with bad performance for adjustment. 2) LSE technique performs better than MLFE which shows that using size measure only is more predictive than using all size related features.

We use the Bonferroni-Dunn test to compare the OABE method against other methods as shown in Figure 3. The plots have been obtained after applying ANOVA test followed by Bonferroni test. The ANOVA test results in p-value close to zero which implies that the differences between two methods are statistically significant based on AR measure. The horizontal axis in these figures corresponds to the average rank of methods based on AR. The dotted vertical lines in the figures indicate the critical difference at the 95% confidence level. Obviously, the OABE methods generated lower AR than other methods over most datasets except for small datasets. For such datasets, all models except NN generated relatively similar estimates but with preference to OABE that has smaller error. This indicates that OABE adjustment method is far less prone to incorrect estimates.

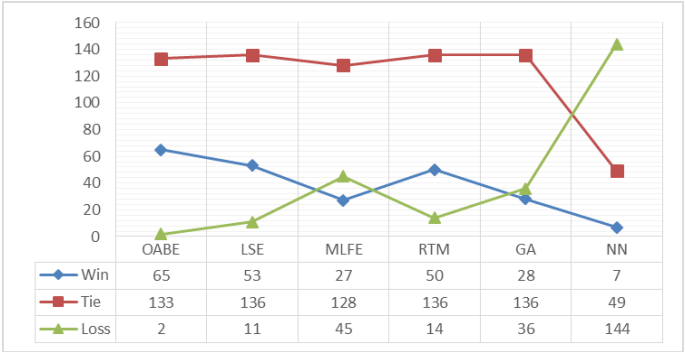


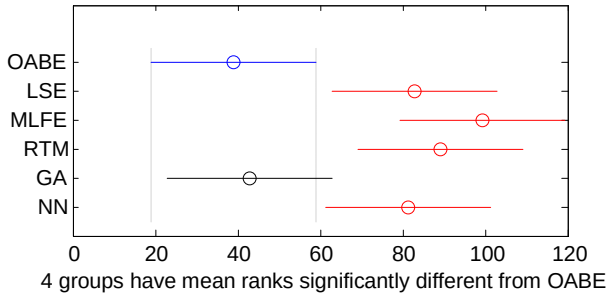
Fig. 2. Win-Tie-Loss Results for all Models

VI. CONCLUSIONS AND FUTURE WORK

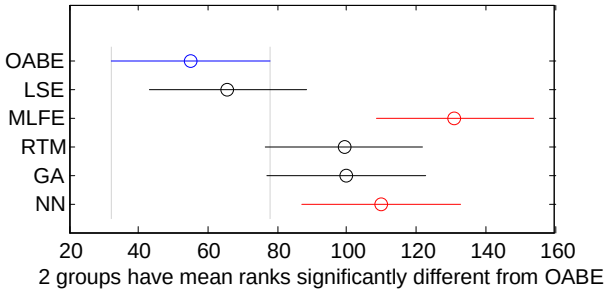
This paper presents a new adjustment technique to tune ABE using Bees optimization algorithm. The BA was used to automatically find the appropriate *k* value and its feature weights in order to adjust the retrieved *k* closest analogies. The results obtained over 8 datasets showed significant improvements on prediction accuracy of ABE. We can notice that all models' ranking can change by some amount but OABE has relatively stable ranking according to all error measure as shown in Figure 2. Future work is planned to study the impact of using ensemble adjustment techniques.

VII. ACKNOWLEDGEMENT

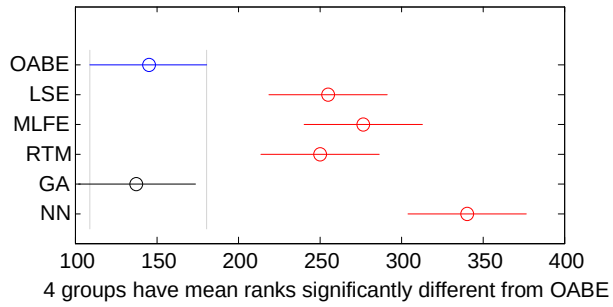
The authors are grateful to the Applied Science Private University, Amman, Jordan, for the financial support granted to this research project.



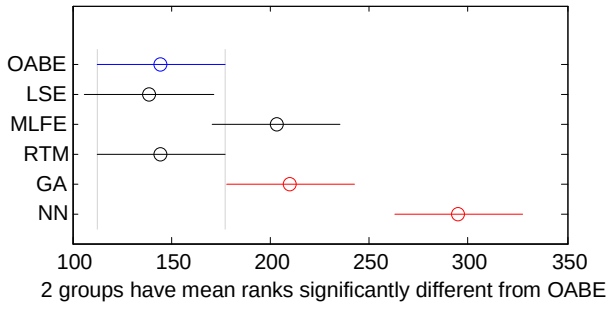
(a) Albrecht dataset



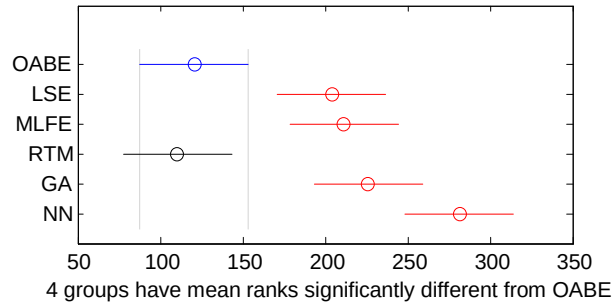
(b) Kemere dataset



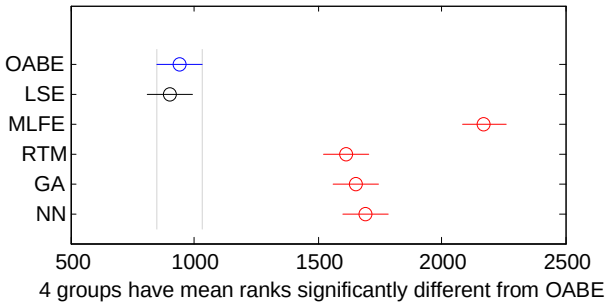
(c) Desharnais dataset



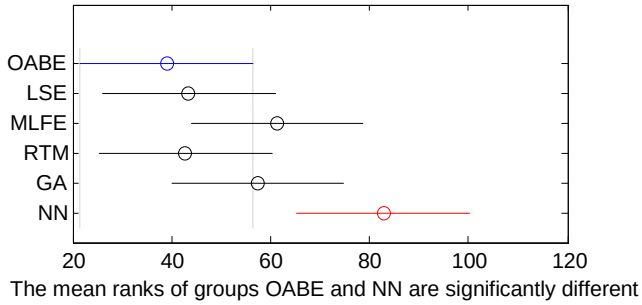
(d) COCOMO dataset



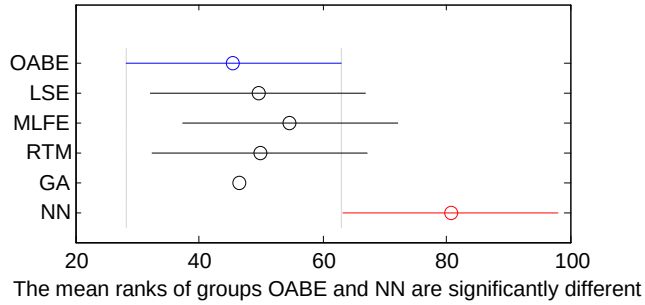
(e) Maxwell dataset



(f) China dataset



(g) Telecom dataset



(h) NASA dataset

Fig. 3. Bonferroni-Dunn test multiple comparison test over all datasets

REFERENCES

[1] M. Azzeh, A replicated assessment and comparison of adaptation techniques for analogy-based effort estimation, Journal of Empirical Software Engineering vol. 17, pp. 90-127, 2012.

[2] M. Azzeh, Y. Elsheikh, Learning Best K analogies from Data Distribution for Case-Based Software Effort Estimation, The Seventh International Conference on Software Engineering Advances (ICSEA 2012), pp. 341-347, 2012.

[3] G. Boetticher, T. Menzies, T. Ostrand, PROMISE Repository of empirical software engineering data http://promisedata.org/ repository, West Virginia University, Department of Computer Science. 2012.

[4] N. H. Chiu, S. J. Huang, The adjusted analogy-based software effort estimation based on similarity distances, Journal of System and Software, Vol. 80, pp. 628–640, 2007.

[5] A. Idri, A. Abran, T. Khoshgoftaar, Fuzzy Analogy: a New Approach for Software Effort Estimation, 11th International Workshop in Software Measurements, pp. 93-101, 2001.

[6] M. Jorgensen, U. Indahl, D. Sjoberg, Software effort estimation by analogy and "regression toward the mean", Journal of System and Software, vol. 68, pp. 253-262, 2003.

[7] C. Kirsopp, E. Mendes, R. Premraj, M. Shepperd, An empirical analysis of linear adaptation techniques for case-based prediction, Internation

conference on Case-Based Reasoning Research and Development, pp. 1064-1064, 2003

[8] E. Kocaguneli, T. Menzies, A. Bener, J. Keung, Exploiting the Essential Assumptions of Analogy-based Effort Estimation, Journal of IEEE transaction on Software Engineering, vol. 38, pp. 425-438, 2011.

[9] J. Z. Li, G. Ruhe, A. Al-Emran, M. Richter, A flexible method for software effort estimation by analogy, Journal of Empirical Software Engineering, vol. 12, pp. 65–106, 2007.

[10] Y. F. Li, M. Xie, T. N. Goh, A study of A study of the non-linear adjustment for analogy based software cost estimation, Journal of Empirical Software Engineering, vol. 14, pp. 603–643, 2009.

[11] U. Lipowezky, Selection of the optimal prototype subset for 1-nn classification, Pattern Recog. Letters, vol. 19, pp. 907–918, 1998.

[12] E. Mendes, N. Mosley, S. Counsell, A replicated assessment of the use of adaptation rules to improve Web cost estimation, International Symposium on Empirical Software Engineering, pp. 100-109, 2003.

[13] I. Myrtveit, E. Stensrud, M. Shepperd, Reliability and validity in comparative studies of software prediction models, Journal of IEEE Transaction on Software Engineering, vol. 3, pp. 380–391, 2005.

[14] D. T. Pham, A. Ghanbarzadeh, E. Koç, S. Otri , S. Rahim, M. Zaidi, The Bees Algorithm – A novel tool for complex optimisation problems, Proceedings of the 2nd Virtual International Conference on Intelligent Production Machines and Systems, pp. 454-459, 2006.

[15] M. Shepperd, C. Schofield, Estimating software project effort using analogies, Journal of IEEE Transaction on Software Engineering, vol. 23, pp. 736–743, 1997.

[16] F. Walkerden, D. R. Jeffery, An empirical study of analogy-based software effort Estimation, Journal of Empircal Software Engineering, vol. 4, pp. 135–158, 1999.

[17] M. Azzeh, "Adjusted case-based software effort estimation using bees optimization algorithm. International conference on Knowlege-Based and Intelligent Information and Engineering Systems. Springer BerlinHeidelberg, pp. 315-324, 2011.

Impediments of Activating E-Learning in Higher Education Institutions in Saudi Arabia

Case Study: Ar-Rass College of Science and Arts, Qassim University

Ashraf M. H. Abdel Gawad
Department of Computer Science
Faculty of Science and Arts Qassim University,
Saudi Arabia

Khalefah A. K. Al-Masaud,
Department of History
Faculty of Arabic Language and Social Studies Qassim University, Saudi Arabia

***Abstract*—This paper presents the real reasons which constraint the application of the E-learning in higher education institutions in Saudi Arabia (Case study: Qassim University)and some suggested solutions. A questionnaire has been designed for the study include 48 paragraphs, divided into 5 parts, the first include the principle information, the second define how the technology can be used in the E-learning, the third deals with how to support the E-learning idea, the fourth part, asking the difficulties and challenges that face the application of E-learning, the fifth items asking to provide suggestion for solving the problem. The study has 100 samples for faculty members and undergraduate students in Ar-Rass college of Science and Arts, male Departments at Qassim University. The study indicates that the main factors that obstruct the E-learning is the financial support from saving advanced PC's, labs, and establishing strong computer network, adding to the weakness of some faculty members and student to English language. The study focused in the suggested solution for the problem by applying the Electronic subjects, and imposes the whole faculty members to prepare at least one course in Electronic form.**

Keywords—E-Learning; Internet; curriculum; Saudi Arabia Universities; cultural aspects

I. INTRODUCTION

Due to penetration of advanced technology, computer networks, and communication technology in all sectors of life makes the applications of E-learning much more easier than before[1-2].

E-learning is considered one of the distance learning forms, and can be defined as a method of using advanced technologies in computer science, such as network, software, multimedia, electronic library, international network (internet), and multimedia. The important is to deliver the information using easiest way in a short time i.e. saving time and effort. As a result of rapid advanced in information technologies and its effect in all aspects of life, the university education must be more responsive for these dramatic changes and rapid development. Due to the fact that the production of the university education is considered an input and work items in different economic sectors, the E- learning need a well-established infra-structure of a computer networks, so that the institutions think about the cost which can sure the success of E-learning process [3].

E-learning introduces a new form of education and getting knowledge around the world. It can help those whom they live in a remote area to continue their university study by application of advanced in technology in computer networks [4].

E-Learning has amazing prospects for the most of the development countries. For successful deployment of e-Learning as a modern teaching method, the readiness of both quantitatively and qualitatively must be measured [5-6].

E-Learning enables the universities and institutions to train their geographically scattered workforce and make them eligible with the dynamic knowledge and skill demands with greater efficiency but at less cost [7-8].

This paper deals with the Impediments of Activating E-Learning in Higher Education Institutions in Saudi Arabia, and how we can overcome the difficulties face the application of E-learning. This can be obtained by a questionnaire filled by student and faculty members in 9 departments at Ar-Rass College of Science and Arts, Qassim University, Saudi Arabia.

II. E-LEARNING DEFINITIONS

It is well known that the E-learning have many definitions, a few of them will be presented here to extract some understanding [9].

1) Tom Kelly, Cisco: "E-learning is about information, communication, education and training. Regardless of how trainers categorize training and education, the learner only wants the skills and knowledge to do a better job or to answer the next question from a customer."

2) "E-learning provides the potential to provide the right information to the right people at the right times and places using the right medium.

3) "*Brandon Hall:* "…instruction that is delivered electronically, in part or wholly via a Web browser, (…) through the Internet or an intranet, or through multimedia platforms such CD-ROM or DVD." Brandon Hall argues that, as the technology improves, e-learning has been identified primarily with using the web, or an intranet's web. Increasingly — as higher bandwidth has become more accessible — it has been identified primarily with using the

Web, or an intranet's web, forcing the visual environment and interactive nature of the web on the learning environment.

4) Learning Circuits: "E-learning covers a wide set of applications and processes such as web-based learning, computer-based learning, virtual classrooms and digital collaboration. It includes the delivery of content via the Internet, intranet/extranet, audio and videotape, satellite broadcast, interactive TV and CD-ROM."

5) Rosenberg: "E-learning refers to the use of Internet technologies to deliver a broad array of solutions that enhance knowledge and performance." Rosenberg claims that e-learning is based on three fundamental criteria:

III. Benefits and Disadvantages

The learning modalities have their strengths and weaknesses. The list of benefits and disadvantages of e-learning can be summarized as the following[10]:

a) Benefits of E-Learning

*1) Reduced cost***.** The total cost can be reduced through reducing the instructor costs, travel expenses, room rentals, lodging and meals. The time required for attending the class could be used for other duties.

2) Efficient. The e-learning is efficiently uses resources to train many people. In synchronous class one instructor can be sufficient for many classes in many locations in the same time. **Globally consistent.** E-learning provides a global solution to the employees whom they work away from the home office.

3) Scalable. E-learning solutions scale more easily than traditional classroom training. **Universal access to experts**. Asynchronous e-learning overcome the problem of communications with the instructor. E-learning provides cross-border access and/or exposure to expert knowledge and top instructors.

*4) Reduces/eliminates travel***.** E-learning eliminates the travel expenses.

5) Trackable. E-learning solutions can provide tracking mechanisms that record attendance, completion and time spent on specific training modules.

6) Convenient. Asynchronous e-learning solutions allow a student to learn based on their personnel circumstances. Synchronous e-learning allows students in different locations to attend classes morning or evening.

b) Disadvantages

1) Initial investment: The costs to establish an e-learning infrastructure can be hit the ledger in Year One, although the e-learning program may have a lifespan that lasts several years.

2) Inappropriate content: The content must match the community requirements. Complex issues that require hands-on learning may not fit the model.

3) Technology issues: As bandwidth and hardware costs continue to decrease, e-learning becomes more relevant as a learning solution. Remote areas with limited bandwidth may not be able to realize the benefits of e-learning.

4) Diminished personal interaction: E-learning limits personal interactions between the instructor and students. The forms of communication are dramatically limited with e-learning. Instructors may find it difficult to determine the level of the students of the subject matter.

5) Employee acceptance: Due to rapid development in communications and technology, more employees became familiar with it, and the acceptance for e-learning grows. Some employees particularly may feel uncomfortable with e-learning. Cultural issues may also inhibit the use of technology for e-learning.

6) Motivation: E-learning, particularly asynchronous training, requires students to take the initiative to start and complete the training. Some students may not be motivated to allocate the time to learn.

IV. The Study Questions:

The study focus on determining the reasons that prevent application of E-learning in Saudi Arabia Higher institutes, and how can overcome the difficulties to spread the E-learning. The problem formulations can be determining through the following questions:

1) Is the current technology is useful in applying E-learning?

2) What are the motivations that support the idea of using E-learning?

3) What are the difficulties that face applying E-learning?

4) What are the suggested solutions to this problem?

V. Importance of The Study

This study aims to find out the difficulties and constraints that face the E-learning in the university community from the student and faculty members' point of view. Also, how to overcome the problem of applying the E-learning in a wide range, regarding the need of community, to this type of education.

VI. Object of The Study

The study aims to the following points:

1) Find out the reasons that obstruct the application of E-learning in the higher education institutes in Saudi Arabia.

2) The feasibility of the current technology in supporting the E-learning.

3) Find the motivation that can support the E-learning.

4) Defining the solution of the problem and how to apply it.

Limitations of the Study:

The study has been developed within the following limits:

1) The factors that effect in e-learning applications in Higher Institutions.

2) Sample from the undergraduate student and Faculty members.

3) Defining the problem through 4 main factors.

4) The questionnaire is applied in the first semester of the academic year 1432/1433 H.

VII. Analysis of The Study:

The questionnaire has been distributed to all departments in the college. The distribution process considering the faculty members belongs to the science departments, the faculty members belongs to Arts department, the student belongs to science departments, and the student belongs to Arts Departments. A table 1 through 4 indicates the results of the questionnaire.

From the results of the study we can summarize the following points:

1) The use of internet chat relay group, E-mail, discussion group and computer based instructions are the most suitable tools that can supporting the use of E-learning.

2) The motivations that support the E-learning process such as provide the chance for completing the university study, helping the student in studying their subjects with less effort, and helping the student to develop their own skills.

3) The Impediments that face the applications of E-learning can be summarized from the study as the following points:

a) Some of the students and some faculty members survive from the poor culture of using PCs.

b) Poor infrastructure for the computer networks.

c) The computer labs in the college is not sufficient to handle the process.

d) Some of the faculty members and student have poor English.

e) There is not enough training program to develop the culture of E-learning.

1) The following are the suggested points to overcome the applying the E-learning process at the college:

a) Providing training program to the student to develop their ability to use the computer.

b) Applying the E-subjects.

c) The faculty members must form at least one of their subject in one of the know E-Form.

d) Providing the internet to the office of the faculty members with advanced PCs.

e) Encourage the faculty members to apply E-learning.

f) Attract the faculty members from those are professional in applying E-learning.

g) Introduce the E-test for some subjects.

h) Establish an individual budget to buy computer programs that help in applying E-learning.

TABLE I. Second Paragraph: Do You Think The Use of The Following Technologies In The Process of E-Learning Is Useful?

	Totally agree	Agree	Neutral	disagree	Totally disagree
1	100%				
2	100%				
3	100%				
4		84%	16%		
5		80%	20%		
6			60%	40%	
7			56%	44%	
8			52%	36%	12%

TABLE II. **Third Paragraph :** Motivations that can support the idea of using E-learning

	Totally agree	Agree	Neutral	disagree	Totally disagree
1	80%	20%			
2	88%	12%			
3	40%	60%			
4	76%	24%			
5	44%	56%			
6	56%	44%			
7	64%	36%			
8	72%	28%			
9	88%	12%			
10	84%	16%			

TABLE III. Fourth Paragraph: Difficulties and Challenges Facing The Application of E-Learning

	Totally agree	Agree	Neutral	disagree	Totally disagree
1		80%	12%	8%	
2		76%			
3			8%	16%	76%
4			12%	72%	16%
5			8%	84%	8%
6				88%	12%
7		84%	8%	8%	
8			4%	8%	88%
9			8%	84%	8%
10	8%	88%	4%		
11			8%	80%	12%
12	4%	84%	12%		
13			4%	92%	4%
14	80%	8%	12%		
15			8%	80%	12%
16			4%	4%	92%
17			4%		96%
18	8%	84%		4%	
19	4%	88%	8%		
20				92%	8%

TABLE IV. FIFTH PARAGRAPH: PROPOSALS THAT CAN CONTRIBUTE TO OVERCOMING THE PROBLEMS OF APPLICATION OF E-LEARNING

	Totally agree	Agree	Neutral	disagree	Totally disagree
1	4%	92%	4%		
2	4%	88%	8%		
3	12%	84%	4%		
4	4%	92%	4%		
5	4%	96%			
6	8%	92%			
7		96%	4%		
8	4%	96%			
9		96%	4%		
10	8%	88%	4%		

VIII. RESULTS AND RECOMMENDATIONS

The study indicated that there are some tools can help in applying the E-learning such as E-mail, internet chat replay group, and computer based instructions. Also, the motivation that can support the idea of E-learning is that helping those whom lives a way from the institutions to complete their university study and develop the student skills. The E-learning provided some services for the student and the faculty members. The major factors that resist the application of E-learning are summarized as the following points:

- Poor culture of using PCs.
- The infrastructures for the computer networks are not well structured.
- The computer labs need to be provided with advanced computer systems.
- low level of English language for Some of the faculty members and students.
- There is not enough training program to develop the culture of E-learning.

In order to remove the resistance of applying E-learning, there are some points must be considered, such as establishing a training program for faculty members and students, applying E-Test for a certain number of subjects, providing the internet service to the whole buildings in the college. The results of the study could be useful for other countries with the same circumstances and culture.

IX. CONCLUSIONS

The paper presents and analysis of factors affecting, the application of E-learning at Qassim university (Ar-Rass College of Science and Arts, Qassim University). Also, presents how can face the problem and improve the culture of e-learning in the university community. The study propose the factors that help in applying e-learning such as E-mail groups, internet chat groups and computer based instructions. Also, applying the E-courses for few subject will force both students and faculty members to raise their ability to deal with E-learning. The difficulties that face applying E-learning is discussed and presented. The result indicate the importance of finding a good media for applying E-learning such as establish adavanced computer laps, establish a training programs to train and the students and faculty members for using E-learning.

X. SUGGESTED FUTURE WORK

The work can be applied for females department to judge if the same difficulties or another factors rather that found in males department. Different factors could be considered at the questionnaire. Reduce the factors of questionnaire in order to determine exactly the difficulties and how to face it within simple applicable procedures.

REFERENCES

[1] Sarah El-Gamal, and Rasha Abd El Aziz, " The Perception of Students Regarding E- Learning Implementation in Egyptian Universities- The Case of Arab Academy for Science and Technology" The 3rd Int. Conf. on Mobile, Hybrid, and On-line Learning, 2011, pp. 1-5.

[2] Abd El Aziz, R. (2009), ATM Location and Usage in Egypt: Social and Technical Perspectives, PhD Thesis, University of the West of England, Brisol, UK.

[3] Wagner, N., Hassanein, K., and Head, M. (2008). Who is responsible for E-Learning Success in Higher Education? A Stakeholders' Analysis. Educational Technology & Society, 11 (3), 26-36. [Online]. (Accessed: 2 December 2010). Available at: http://www.ifets.info/journals/11_3/3.pdf.

[4] Nagwa El Shenawi, "E-Learning, Challenges and Opportunities: The Case of Egypt" Egypt, 2004.

[5] Chandan Kumar Karmakar, CM Mufassil Wahid, RECOMMENDATIONS FOR BANGLADESH TOWARDS E-LEARNING READINESS" Proceedings of the 2nd Int. Conference on eLearning for Knowledge-Based Society, August 4-7, 2005, Bangkok, Thailand.

[6] Spiros Ap. Borotis and Angeliki Poulymenkou.] "E-Learning Readiness Components: Key Issues to Consider Before Adopting e-Learning Interventions" The leading digital library dedicated to education and information technology, Paper Id:11555, found at: http://editlib.org/noaccess/11555.

[7] Rosenberg, M.J. (2000a). e-Learning: Strategies for Delivering Knowledge in the Digital Age: McGraw-Hill.

[8] Rosenberg, M.J. (2000b). The E-Learning Readiness Survey. Retrieved February 2004, from: http://ww.books.mcgrawhill.com/training/elearning/eLearning_Survey.pdf

[9] E-learning, A research note by Namahn, it found in: www.namahn.com/resources/.../note-e-learning.pdf

[10] The Value of E-learning, IBM Training, White Paper: found at: http://www.ibm.com/software/lotus/training.

XI. APPENDIX: (QUESTIONNAIRE USED FOR THE STUDY) NAME(OPTION):

NAME(OPTION):

FIRST: BASIC INFORMATION: * PLEASE TICK THE APPROPRIATE BOX THAT APPROPRIATE TO YOUR STATUS.			
1-GENDER		2- PROFESSION	
FEMALE ( )	MALE ( )	FACULTY MEMBER ( )	STUDENT ( )
3- ACADEMIC QUALIFICATIONS		4- DEPARTMENT	
PH.D. ( )	M.SC. ( )	COMPUTER SCIENCE ( )	PHYSICS ( )
		ISLAMIC STUDY& QURAAN ( )	CHEMISTRY ( )
B.SC. ( )	OTHER ( )	ENGLISH LANGUAGE ( )	SPECIAL EDUCATION ( )
		ARABIC LANGUAGE ( )	BASIC EDUCATION ( )
		OTHER ( )	
5- **LEVEL OF PROFICIENCY IN DEALING WITH THE COMPUTER AND APPLIED SOFTWARE**		6- DO YOU USE THE PC IN YOUR WORK (EDUCATION)	
BEGINNER ( )	GOOD ( )	YES ( )	NO ( )
VERY GOOD ( )	PROFESSIONAL ( )		
7- DO YOU HAVE PC IN YOUR HOME (OFFICE)?		8- HOW MANY TIMES YOU USE INTERNET?	
YES ( )	NO ( )	DAILY ( )	ONCE EACH 2 DAYS ( )
		ONCE EACH 3 DAYS ( )	RARELY ( )

Second: What do you think the use of the following technologies in the process of e-learning * Please tick the appropriate box of the extent of your agreement to these justifications.		**Totally agree**	**agree**	**Neutral**	**disagree**	**Totally disagree**
1	Electronic Mail					
2	Internet					
3	Internet Relay Chat (IRC)					
4	Discussion Group					
5	Computer Based Instruction					
6	Video Conference					
7	Virtual Class					
8	News Group					

Third: Motivations that can support the idea of using e-learning ***** Please tick the appropriate box of the extent of your agreement to these justifications.		**Totally agree**	**agree**	**Neutral**	**disagree**	**Totally disagree**
1	Given many opportunities to pursue higher education					
2	Help students to communicate with the college easily					
3	Help students complete their college education and to overcome the geographical difficulties.					
4	Make communication between students and faculty members in the process very easy.					
5	Help students to communicate with each other.					
6	Help develop students' skills in dealing with the computer					
7	Help in the preparation of qualified human resources, which are commensurate with the quality programs.					
8	Help create a competitive environment between the educational institutions					
9	Helps to spread higher education in areas that do not have institutes or colleges.					
10	Help to provide electronic services to students, as well as faculty members.					

Fourth: Difficulties and challenges facing the application of e-learning *** Please tick the appropriate box of the extent of your agreement to these justifications.**		**Totally agree**	**agree**	**Neutral**	**disagree**	**totally disagree**
1	Lack of students for the culture of dealing with the computer					
2	Lack of some members of the faculty of the culture of dealing with the computer.					
3	Not convinced some members of the faculty of the feasibility of using e-learning.					
4	Convinced some faculty members that this type of education does not bear fruit.					
5	Lack of awareness of the importance of overall management of this type of education.					
6	Not convinced of the overall management of the importance of using this type of education					
7	The unwillingness of students to use this type of education.					
8	Difficult to provide Internet services.					
9	Not to encourage faculty to use this type of education.					
10	Lack of modern computers will help spread this type of education					
11	The lack of computers among students in their homes					
12	Poor infrastructure necessary for the work units of this type of education.					
13	Lack of financial resources needed to build a good network computer.					
14	Lack the necessary training programs to promote this type of education					
15	The unwillingness of the faculty members in training programs that help them to master this type of education.					
16	The lack of computers in many of the faculty members in their homes					
17	The lack of computers in many of the faculty members in their offices.					
18	Poor English language proficiency among students.					
19	Poor English language proficiency among some faculty members.					
20	Lack of community acceptance for this type of education.					

Fifth: Proposals that can contribute to overcoming the problems of application of e-learning * Please tick the appropriate box of the extent of your agreement to these justifications.		**Totally agree**	**agree**	**Neutral**	**disagree**	**Totally disagree**
1	Rehabilitation of the student to deal with the establishment of computer training programs.					
2	The application of e-courses effectively.					
3	Establishment of training courses for students and faculty members to enable them to implement e-learning.					
4	Requiring members of the faculty composition of decision-mail at least those that they teach.					
5	Provide a computer for each member of the faculty office.					
6	Make the internet service accessible within the college.					
7	Motivate and encourage faculty members to implement e-learning.					
8	Attract qualified human resources to implement this type of education.					
9	Introduce of electronic tests on some courses					
10	A separate budget for the purchase of software that contribute to the implementation of e-learning.					

Dynamic Allocation of Abundant Data Along Update Sub-Cycles to Support Update Transactions in Wireless Broadcasting

Ahmad al-Qerem
Department of computer science,
Zarqa University Zarqa, Jordan

***Abstract*—Supporting transactions processing over wireless broadcasting environment has attracted a considerable amount of research in a mobile computing system. To allow more than one conflicting transactions to be committed within the same broadcast cycle, the main broadcast cycle need to be decompose into a sub cycles. This decomposition contains both the original data to be broadcast in Rcast cycle and the updates come from the committed transactions on these data items called Ucast cycle. Allocation of updated data items along Ucast cycles is highly affecting the concurrency among conflicting transactions. Given the conflicting degree of data items, one can design proper data allocation along Ucast Cycles to increase the concurrency among conflicting transactions. We explore in this paper the problem of adjusting abundant data allocations to respond in effective way to the changes of data conflicting probability, and develop an efficient algorithm ADDUcast to solve this problem. Performance of our adjustment algorithms is analyzed and a system simulator is developed to validate our results. It is shown by our results that the ADDUcast is highly increased the average number of committed transactions within the same broadcast cycle.**

Keywords—data broadcast; Dynamic Allocation; concurrency control; cycle decomposition; Abundant Data

I. INTRODUCTION

Data broadcast is becoming a promising way to disseminate information to a large population of mobile clients by mean of transaction. Unlike the conventional client server approach, where a data item have to be send many times to deliver the requested data even in the case of read –only transactions. Broadcast has the potential to satisfy all outstanding requests for the same data with a single response. It increases the efficiency of shared bandwidth and improves the system throughput. However, existing mobile technologies have to face several constraints such as limited network bandwidth, frequent disconnections and insufficient battery power. To cope with these constraints, there has been many studies on data transmission techniques using wireless data broadcasting [1], [7] and [8]. Generally a mobile client sends requests to the server and receives the response. For sending requests, mobile clients have to consume uplink bandwidth. The response time also can dramatically increase when the server is heavily loaded by the requests from a large number of clients. However, wireless data broadcast models can overcome these problems. For example, [1] proposed the broadcast disks model. The server continuously and repeatedly broadcasts all data in the database using a single or multiple wireless communication channels. Clients wait for the data in need to come up on the channel, and retrieve data from the channel. In this system, the number of mobile clients does not affect their access time. With its good scalability, wireless data broadcast is used in various mobile applications, e.g. auctions, electronic bidding, stock trading, weather information and traffic information broadcasts [9]. In these applications, the consistency among data items is likely to be violated by update transactions [8], [10] and [16]. Thus, a concurrency control scheme is needed to preserve data currency and consistency for mobile transactions. However, conventional concurrency control methods cannot be directly applied to mobile transaction processing [10], [11]. On the other hand, when an application involves wireless mobile clients that run multiple-operation transactions and dynamically update the server database, those updates have to be appear in the next broadcast cycle. Earlier show up of such updates is highly improving the performance. In this paper we aim to decompose the main broadcast cycle into sub cycles which only contain a subset of data items in the database. This decomposition contains both the original data to be broadcast and the updates come from the committed transactions on these data items. At sub cycle level, it is more powerful for both update and read-only transactions which allow more than one conflicting transactions to be committed on the same broadcast cycle. Moreover, by using sub cycle the currency of the data may be higher since the delays in performing the updates are shorter. There have been many research efforts reported in the literature that tackle the concurrency problems in wireless broadcast environments, such as Update First Ordering (UFO) [5], Multi-Version Broadcast [2, 3], Serialization Graph [2, 3], Broadcast Concurrency Control with Time Stamp Interval (BCC-TI) [6], F-Matrix [4], and Certification Report [7]. The drawbacks of these methods have been analyzed in [12, 13]. In general, some of these methods only support client read-only transactions, and some of them could have substantial processing overhead.

The major contribution of this paper is determining the proper allocation of abundant updated data item along the remaining Ucast cycles and dynamically adjustment of such allocation to support update transactions responding in effective way to the changes of data conflicting probability, and develop an efficient algorithm DADUcast to solve this problem. According to the extensive analysis and comprehensive performance evaluation, the proposed approach shows a satisfactory performance in transactions processing on

broadcast environments. It is shown by our results that the DADUcast is highly increased the average number of committed transactions within the same broadcast cycle.

The rest of this paper is organized as follows: Section II describes the system architecture. Section III illustrates the problem of abundant data and how it could affect the transaction processing. Section IV proposes the ADDUcast approach in more details. Section V presents the performance evaluation of our approach and show the superiority of the proposed algorithm. Finally, we conclude the paper in Section IV.

II. System Model

The system model of adopted in this paper allows both read-only transactions (reading from air without being known by the server) and update transactions (modifying data by sending requests through a low bandwidth channel after completion at client). All operations are executed in order.

The broadcast cycle is divided into multiple sub-cycles of two types alternatively see Fig.1: read cast called Rcast and update cast called Ucast in alternative way. The former contains the data items which were predefined scheduled for the main broadcast cycle but distributed along many Rcast whereas the Ucast content is dynamics and changed based on the data being updated by the committed transactions in the previous Rcast cycles. Between any two Rcast cycles, there is a reserved space for Ucast to accommodate identities for all the data objects which are updated by transactions in the server after the first sub-cycle begins.

A mobile transaction can validate its prefetched data consistency autonomously by accessing the Ucast cycle. Our goal is to reveal the write set of the committed transactions as soon as possible. Unfortunately, this may jumble the original schedule. We can tradeoff the immediate show up of the updated data items from the previous sub cycle and delaying all those updates until the beginning of a next sub cycle which will be accommodated in front of pre assigned Rcast index. As a result, the identities of all the data items which are updated by any transactions are included in the Ucast space after the current broadcast sub-cycle.

The identities of all the extra data items with no rooms at the current UCAST are included in the next UCAST based on their conflicting probability. At the same time, since all information of latest committed update transactions is dynamically loaded at the beginning of some Rcast index segment, a mobile transaction only needs to tune to the channel at specific periods to retrieve the requested data item and the control information associated with it.

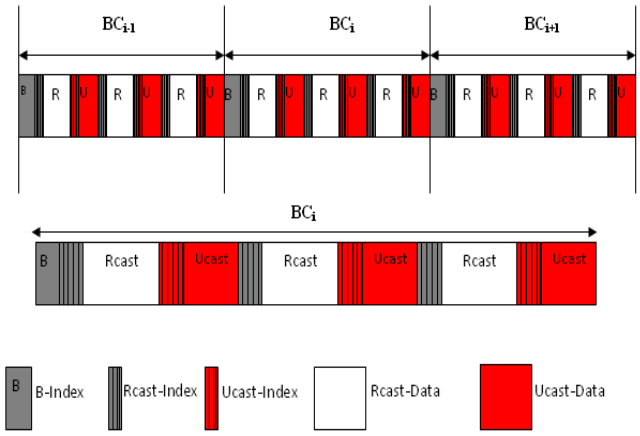


Fig. 1. Broadcast Cycle Decomposition

Data broadcast usually requires a client to be active all the time in order to monitor the data units that go by. This leads to unacceptable energy consumption on wireless mobile equipments, for which power saving is a very essential issue. To save power in data broadcast models, indexing schemes are proposed in [17]. The Basic idea of indexing is to insert pointers for data broadcast in a future schedule into a broadcast cycle. Consequently, a client application can go to doze mode after it accesses this pointer, and only wakes up at the time the requested data unit is on the air. Several index schemes have been proposed in. In all indexing schemes, an index tree of all data in a broadcast cycle is inserted to the schedule. Pointers to each real data units are located at the leaves of the index tree while a route to a specific leaf can be found following the pointer from the tree root. In our work the B-index contains only information about the starting time of both Rcast-index and Ucast-index based on the number of sub cycles and the sizes for each of them; this is easy know as all Ucast are of equal size to and the Rcast cycles is previously determined at the beginning of each broadcast cycle. In addition to the index information provided by Rcast-index and Ucast-index, Rcast-index contains a pointer for the next Rcast. the Ucast-Index contains the actual index of the data items to be broadcasted and each index is associated with control information to validate its pre fetched data items such as sub cycle number as well as the conflicting degree which will be used in allocation and adjustment procedure as we will be explain later in this paper. Fig.2 shows the structure of index information for each element in Ucast-index in addition to the pointer indicating the next Rcast-index segment as the client may arrive at any time during the major broadcast cycle.

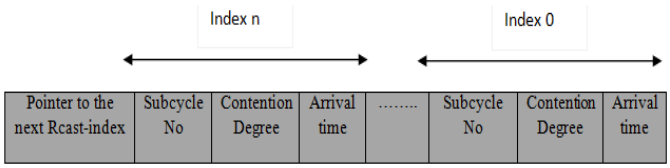


Fig. 2. Information associated with each index in Ucast cycle

III. PROBLEM DESCRIPTION

By exploiting the feature of tree generation with variant-fan-out [14], [15] a heuristic algorithm to distribute the abundant data items along the coming Ucast cycle is developed. The tree obtained in [14] is called Ucast allocation tree where the depth of the allocation trees corresponds to the number of Ucast cycles, and those leaf nodes in the same level of the allocation tree correspond to those data items to be put in the same Ucast cycle. Fig 3 shows a hierarchical broadcast program with two random Ucast allocation tree where the upper level corresponding to the current Ucast cycle is allocated with two data items(assuming the Ucast size =2) and each other lower levels is distributed along the next coming Ucast2 ,Ucast3 respectively**.** As such, the data items in the upper level must be accommodated first satisfy more transactions than those data items in the lower level.

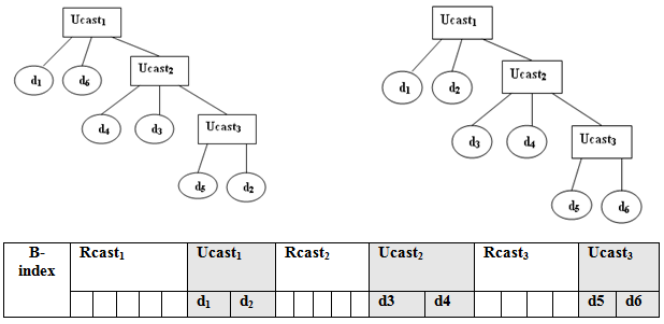


Fig. 3. Updated data item during **$Rcast_1$** = **{d1, d2, d3, d4, d5, d6}** and the size of **$Ucast_1$** = **2**. The abundant updated data items {**d_3, d_4, d_5, d_6**} need to be distributed a long the next coming Ucast cycles to maximize the number of committed transaction within one broadcast cycle.

Note, however, that the algorithm in [14] is designed for the situation where the contention degree of the data items being updated determined at the beginning of broadcast cycle. This consider non practical, as the transactions who wait for this data may need to span multiple broadcast cycle in order to have an opportunity to finish their execution and commit. Clearly, without adapting to the change of update frequencies, the broadcast program determined off-line will unavoidably lead to degraded performance. Thus, with the broadcast programs generated by [14], it is important for the broadcast programs to dynamically adapt to the change of the update frequencies during broadcast cycle so as to retain the performance and increase the freshness of the data items for both read-only and update transactions.

The problem we study can be best understood by the illustrative example in Fig.4. Assume that the data items d_i, $1 \leq i \leq 6$ are of the same size and the number of transaction validating at the server during $Rcast_i$ and $Rcast_{i+1}$ are 5. Denote that the conflicting degree of data item di as $Cd(d_i)$ which is representing the number of validating transactions in which d_i exists in its write set.

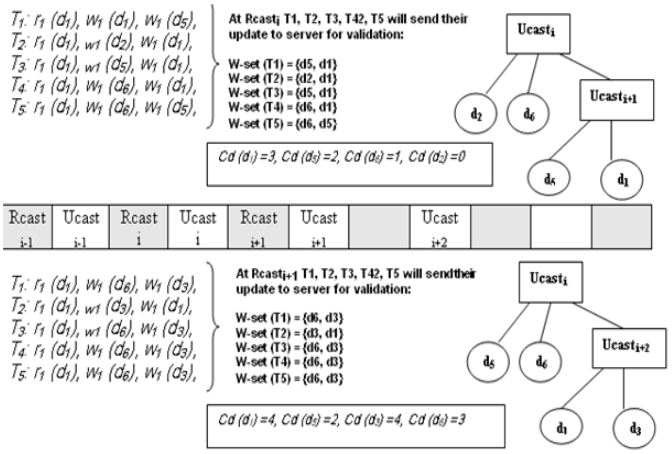


Fig. 4. Illustrative example

Denote the total number of data items being updated in a certain Rcast cycle as n, and a data item as d_i, $1 \leq i \leq n$. The number of Ucast Cycle in a broadcast cycle is S. Recall that $CR(d_k)$ is the conflict ratio of d_k in $Rcast_i$ and estimated as

$$CR(d_k) = \frac{Cd(d_k)}{\sum_{i=1}^{n} Cd(d_i)} \quad \text{.....(1)}$$

Multiplying the conflicting ratio of each data item d_k $CR(d_k)$ by the expected cost of restarting transactions due to conflicting ratio of that data item and summing up the results Same as in [1][19], the expected restarting cost Rc for each data item in the Ucast i is formulated as:

$$Rc = \sum_{x=1}^{N_i} \frac{N_i - x}{N_i} \quad \text{.....(2)}$$

Where Ni us the number of data item allocated in $Ucast_i$. Suppose that the number of updated data items in $Rcast_v$ has j data items, where $j > |Ucast_V|$ and $Uacst_v$ accommodate $d_1, d_2, .. d_{i-1}$, then $d_i, d_{i+1}, ..., d_j$ .. That mean, the updated data items in a $Racst_v$ cannot be accommodated in the corresponding Ucast cycle and have to be transferred to the next Ucast cycle. Unfortunately, the conflicting degree of the updated data items in a certain Rcast with no room in the corresponding Ucast is dynamically changed because of transactions who fetched those data items in Rcasti and finish its execution at Rcast k where k>i. The accumulation of the abundant updated data items come from the latest Rcast cycle need to be distributed properly over the coming Ucast tacking into considerations a dynamic change of their conflicting degree. The conflicting cost *C* of $Ucast_v$ in an allocation tree that has (j-i-1) abundant data items $d_i, d_{i+1} \ldots d_j$ is defined as:

$$C_{i,j} = \sum_{k=1}^{j-i-1} \frac{(j-i-1)-k}{j-i-1} \sum_{\ell=i}^{j} CR(d_\ell) \quad \text{.....(3)}$$

The conflict ratio CR (d_i) is adjusted during Rcast cycles while transactions execution. In essence, the value of $C_{i,j}$ is related to the average restarting cost resulting from the

conflicting degree of abundant data items in Ucast v. Theoretically, generating a broadcast program can be viewed as a partition problem for data items. Given the number of Ucast in a broadcast cycle and the contention degree all data items in the corresponding Rcast, we shall determine the proper set of data items that should be allocated to each Ucast with the purpose of maximizing the number satisfactory request within one broadcast cycle. The problem of distribution updated data items over a K Ucast cycles can be viewed as a discrete maximization problem: Given a list of n data items with their contention degree, partition them into K parts so that the number satisfactory requests of all data items is maximized. As pointed out in previous section, a distribution updated data items over a K Ucast cycles can be represented as an allocation tree with a height of K. Note that the leaf nodes in the same level of the allocation tree correspond to a set of data items to be put in the same Ucast cycle. The value of $C_{i,j}$ is related to the average delay result from the conflicting degree of leaf nodes in level v. This paper investigates the problem of adjusting the data broadcast program at a certain points (i.e. Ucast cycle) to satisfy as much transactions as possible. In order not to distract readers from the main theme of this paper for dynamically adjusting broadcast programs, readers interested in the details of update transactions processing in broadcast data model are referred to [13][16]. Once the change of conflicting degree is larger than the predetermined value, algorithm ADDUcast will be executed to reach the new configuration with minimal conflicting. In accordance with the conflicting degree of data items at Rcast1 and the number of Ucast cycles given, the allocation tree was determined based on the their committed time. It can be seen in Fig 4, at time Rcast $_{i+1}$ allocation tree differ from the one at Rcast $_i$ due to the change of conflicting ratio. Consequently, data items should be moved among levels within the given allocation tree in response to the change of conflicting ratio of data items. Clearly, such movements have an impact on the average response time for all transactions.

IV. Allocation and Adjustments Approach

We devise in this paper an algorithm, referred to as algorithm ADDUcast, to dynamically adjust the broadcast programs by shuffling data items among different levels in the allocation tree. in this paper we propose an algorithm, referred to as algorithm ADDUcast, to dynamically adjust the abundant data allocations by shuffling data items among different Ucasts in order to reflect the contention status therapy increasing the currency and utility of those data items. The process of algorithm ADDUcast can be decomposed into two phases, namely (1) the basement adjustment phase and (2) the smooth adjustment phase. In the basement adjustment phase, algorithm ADDUcast moves data items among the remaining Ucast cycles so as to enable the costs of most Ucasts in the allocation tree to be smaller than or equal to average cost as we describe in algorthim1. Then, for smooth adjustment, algorithm ADDUcast adjusts the data items between consecutives Ucast with the objective of minimizing the total cost of these two consecutive Ucasts.

Since the basement adjustment intends to let the total cost of allocation tree be evenly allocated to all levels, it is possible that some data nodes would move back and forth between neighboring levels.

For execution efficiency, the number of runs for the basement adjustment is limited to be K-1. Basement tuning described in *algorithm 3* is developed to move data items in Ucast i so as to satisfy the purpose of the basement adjustment. By exploiting the basement adjustment, data items are roughly allocated to each Ucast of an allocation tree with the costs of most Ucasts are smaller than or equal to average cost.

The Ucast status table UST is created to record the cost of each Ucast in the allocation tree, and the number of rows in UST is equal to the number of Ucasts in a broadcast cycle (see *algorithm 1* line *7-10*). UST(i).D = UST(i).C −UST(i).P. Where UST(i).P is the cost of data items in Ucast i previously, whereas the value of UST(i).C represent the cost of data items in Ucast i regarding the conflicting probability of the latest Rcast cycle. Also, UST(i).check is a Boolean variable used to indicate whether the basement tuning is performed or not.

1. Algorithm 1: Abundant data Distribution over Ucast cycle ADDUcast
2. Input:
3. K: number of remaining Ucast cycles
4. Usize: number of room in each Ucast cycle.
5. UST: The Ucast status table (UST) with K rows.
6. Output: Minimal conflict Ucast allocation tree up to next Broadcast Cycle
7. /* Construction and initialization of UST content */
8. **for** each row $_i$ in UST **do**
9. **{** UST(i).D=UST(i).C-UST(i).P;
10. UST(i).Checked=false **}**
11. /* Vector Δ has K-1 elements which record the conflicting cost difference between two consecutive Ucasts */
12. For each element i in vector Δ **do**
13. **{** Δ [i]:= |UST(i).C − UST(i + 1).C|
14. Base-checking=0 **}**
15. /* The Basement adjustment phase */
16. Select $r_i \in$ UST such that UST(i).D is maximal;/*Select row of maximal difference */
17. repeat
18. begin
19. **if** (i==1)
20. call Basement Tuning (i, i+1);
21. **else if** (i==k)
22. call Basement Tuning (i, i-1);
23. **else**
24. **{**select the row j where j∈ (i − 1, i + 1) such that ST(j).G is false and Δ [j] is maximal;
25. Call Basement Tuning (i, j); **}**
26. base-checking ++;
27. update UST and Vector Δ accordingly ;
28. select the row i from UST where UST(i).C is maximal and UST(i).Check is false;
29. end
30. **until** base-checking ==K-1;
31. Call smooth adjustment phase

Vector Δ has K-1 elements that record the cost difference between two consecutive Ucast cycles (see *algorithm 1 from line 12 to line 14*). As can be seen in basement adjustment algorthim1, algorithm ADDUcast makes sure that most Ucasts of the allocation tree meet the requirement of the basement adjustment. Since the casual adjustment intends to let the total

cost of allocation tree be evenly allocated to all levels, it is possible that some data nodes would move back and forth between neighboring levels. By take advantage of basement adjustment, data items are roughly allocated to each Ucast of based on corresponding allocation tree with the costs of most Ucasts are smaller than or equal to average cost. For execution efficiency, the number of runs for the basement adjustment is limited to be K-1. Procedure basement tuning is developed to move data items in Ucast i so as to satisfy the purpose of the basement adjustment.

```
1.  Construct a priority queue PQ;
2.  /* A priority queue returns the element with the minimal
    conflicting value */
3.  for each element i in Vector Δ do
4.  Insert Δ [i] into the PQ;
5.  While (PQ is not empty) and Ucast.CurrentSize < USize
6.  begin
7.  remove the element i from PQ;
8.  if (UST(i).C<UST(i+1).C)
9.  /* if there is no movement between Ucast i and Ucast i+1,
    moving equals to -1 */
10. Moving=push up (i, i+1);
11. CurrentSize = CurrentSize +1
12. else
13. Moving=pull down(i, i+1);
14. CurrentSize = CurrentSize -1
15. if (the movement occur: moving <> -1)
16. Update the elements in PQ and UST accordingly;
17. end
```

Then, algorithm ADDUcast employs the smooth adjustment *algorithm 2* to adjust data items between consecutive Ucasts. As can be seen from *line 2 to line 18 of algorithm 2*, consecutive Ucasts are examined on finding potential movements with the purpose of minimizing the total cost of consecutive Ucasts. Specifically, in *line 8 of algorithm 2*, the sequence of performing the smooth tuning is determined by identifying the largest cost difference among those between consecutive Ucasts (i.e., the largest value in Δ). After identifying the consecutive Ucasts (e.g., level I and level i+1) to perform the smooth tuning, one should determine the data movements between these Ucast. Note that there are two kinds of movements, i.e., pushing up and polling down. Judiciously applying these movements is able to reduce the total cost of these two consecutive Ucasts. Clearly, if the cost of Ucast i is smaller than Ucast i+1, we should move data items from level i+1 to level i and vice versa.

From *line 7 to line 18, algorithm 2* adjusts data items in consecutive Ucasts iteratively with the objective of minimizing the total cost until there is no further adjustment required (i.e., queue PQ is empty).

```
1.  Algorithm 3: Basement Tuning (Ucast i, Ucast j)
2.  { sort those data items in Ucast i according their conflict
    probabilities;
3.  if (i<j)
4.  begin
5.  while UST(i).C > Σ_{i=1}^{k} UST(i).C / k do
6.  Move the data item in the rightist side of Ucast i to Ucast j
    and update ST(i).C accordingly;
7.  end
8.  else
9.  begin
10. while UST(i).C > Σ_{i=1}^{k} UST(i).C / k do
11. move the data item in the leftist side of Ucast i to Ucast j
    and update ST(i).C accordingly;
12. end
13. }
```

Once determining the movement's direction among Ucasts, we should determine the number of data items considering the available space in hosting Ucast. Such a number determine based on the conflict reduction gain result from the movement which is limited also by the available space. To do so, we will use the move up and move down procedures as follows: Suppose that data items in each Ucasts are sorted according to the descending order of conflicting probability.

Conflict reduction gain occurs by moving N data items from $Ucast_i$ to $Ucast_{i+1}$ can be estimated by: *CRG-U*(N) = $(C_{i,k} + C_{k+1,j}) - (C_{i,k+p} + C_{k+p+1,j})$ assuming that $Ucast_i$ has k−i+1 data items, d_i, d_{i+1}, ..., d_k and Ucast i+1 has j−k data items, d_{k+1}, d_{k+2}, ..., d_j . The following procedure move-up Fig.5 determine the set of data items in $Ucast_{i+1}$ to be moved upward to $Ucast_i$ in order to maximize the conflict reduction gain between these consecutive Ucasts.

```
Procedure Move-up (Ucasti, Ucasti+1)
{Determine N' such that CRG-U (N') = max{1≤N≤j−k {CRG-U
(N)};
/* determine the maximal value of CRG-U (N)) such that N 1≤N≤ j-k.
*/
if CRG-U (N') > 0
Move data items d_{k+1}, d_{k+2}, ..., d_{k+N"} to Ucasti;
else
N' =-1; /* no movement is performed since there is no cost-effective
movement. */ }
```

Fig. 5. Move-up procedure

In the same way of Move-up procedure, the Move-down procedure in Fig.6 evaluate the set of data items in Ucast i to be moved downward to $Ucast_{i+1}$with the purpose of maximizing the conflict reduction gain of these consecutive Ucasts.

The *CRG-U*(N) for Move-down estimated as = $(C_{i,k} + C_{k+1,j}) - (C_{i,k-p} + C_{k-p+1,j})$ assuming that $Ucast_i$ has k−i+1 data items, d_i, d_{i+1}, ..., d_k and $Ucast_{i+1}$ has j−k data items, d_{k+1}, d_{k+2}, ..., d_j .

```
Procedure Move-down (Ucast_i, Ucast_{i+1})
{Determine N' such that CRG-U (N') = max {1≤N≤ k-i-1 {CRG-
U (N)};
/* determine the maximal value of CRG-U (N)) such that N 1≤N≤ k-
i-1. */
if CRG-U (N') > 0
Move data items d_{k−N"+1}, d_{k−N"+2}, ..., d_k to Ucast_{i+1};
```

else
N' =-1; /* no movement is performed since there is no cost-effective movement. */
}

Fig. 6. Move-down procedure

V. PERFORMANCE EVALUATION

We will investigate the performance of our approach in average response time of the transactions as well as the average restart time of the transactions under different workload and system setting. Table 1 shows the important configuration parameters and their definitions. Read-only and update transactions are simulated based on the ratios defined in the table. Each transaction in the simulation consists of several data access operations and computation operations before or after each data access. A transaction's length is the number of access operations in it. The transaction length is uniformly distributed.

The computation time between two access operations is the operation inter-arrival time. Time between the starts of two consecutive transactions in a simulated system is denoted as transaction inter-arrival time. Each simulation run uses only one client. Large numbers of clients are emulated by using a small average transaction inter-arrival time from one client. Since the amount of abundant data depends on the probability of data conflicts among mobile transactions, we will test our approach when the mobile transactions have different data access patterns. The Zipf distribution with a parameter theta is often used to model non uniform access. It produces access patterns that become increasingly skewed as theta increases. The updates come from the mobile transactions are generated following a Zipf distribution similar to the read access distribution at the client. The update distribution is across the range 1 to UpdateRange. in order to simplify the model, A flat broadcast disk is assumed for selecting data items for broadcast along Rcast cycles.

TABLE I. SIMULATION PARAMETERES

Parameters	Value
decomposition factor	2, 4, 6, 8, 10
Zipf parameter ɵ	0.0–1.0
Database size	10000 items
Ucast size /Rcast Ratio	1/4
Ucast size	20 data items
Transaction size (number of operations)	8,10,12,16
Read operation ratio (for update transactions)	0.5
Read-only transaction ratio	0.7
Average delay between operations	1 (exponentially distributed)
Average delay between transactions	5,10,50 (exponentially distributed)

The number of Ucast and Rcast cycle determined based on the decomposition factor which represent number of Rcast and Ucast cycle in the main broadcast cycle (i.e. when the decomposition factor equal to 2 this mean that the broadcast cycle consists of 2 Ucast and 2 Rcast cycles). Recall that the alternative to the use of dynamic allocation of an abundant data items is to randomly distribute the abundant data along remaining Ucast. A scheme that randomly distributes data items over Ucast cycles, referred to as UD-RAN, is implemented for comparison purposes. To see how the protocols influence the performance, for each configuration the average response time the average restart time are also recorded for UD-RAN and compared with the value obtained under our approach.

Performance of algorithms ADDUcast and UD-RAN is comparatively evaluated It is shown that ADDUcast significantly outperforms RAN due to the proper allocation of data along Ucast that maximize the utilization of data for the active mobile transactions based on the conflict cost heuristic. It is important to see that the advantage of ADDUcast over UD-RAN becomes even more prominent when the skew of updates increases, showing the dynamic allocation and adjustment is even more beneficial when the update is more skewed. See Fig. 7, where the y-axis corresponds to the average response time and the x-axis corresponds to the value of skew factor θ.

In addition, performance of algorithms ADDUcast and UD-RAN is evaluated for different Ucast cycle size. The corresponding results are shown in Fig. 8, where it can be seen that algorithm ADDUcast consistently outperforms algorithm UD-RAN for various values of n. The advantage of ADDUcast over UD-RAN increases as the Ucast cycle size decreases while skew factor is constant. When the size of Ucast is large enough to accommodate all the updated data (i.e.no abndunt data) the performance of ADDUcast are very simller toUD-RAN see Fig. 9.

Fig.7 and Fig.8 show the performance of average response time compared to transaction arrival time among simulated transactions under both UD-RAN and ADDUcast. Each simulation run records the response time of all transactions, which is the time period between a transaction's invoked time and the time it commits. It can include the duration of multiple runs of a transaction if the transaction has ever been aborted and restarted. Average response time is the average response value among all transactions read-only and update transactions without considering the transaction lengths. These figures show the results under different decomposition factor with different Ucast size. It shows that the response time under ADDUcast is less than the UD-RAN as the decomposition factor increase and Ucast size decreases. It also shows that at small Ucast size, the abundant data increase and the allocation of these data items a long remaining Ucast cycles according to ADDUcast is highly decrease the response time of transactions involving in accessing those data items. Actually this is reasonable because the response time is highly affected by the restart rate of transactions as we can see in the next experiments Fig.10, Fig. 11

VI. CONCLUSIONS

The presence of update transactions in wireless broadcast environment make it more difficult to deal with conflicting transactions within the same broadcast cycle. Many researchers tackle the problem of concurrent executions of these

transactions and many of them suggest that broadcast cycle decomposition is a proper technique to increase the concurrency and the system performance as well. As a result, the identities of all the data items which are updated by any transactions in a certain sub cycle we call it Rcast are included in the reserved space after the current broadcast sub-cycle called Ucast. The identities of all the extra data items with no rooms at the current Ucast are included in the next Ucast based on commit sequence of their transactions. Actually, The accumulation of such abundant updated data items is highly affect the system performance and need to be distributed properly over the coming Ucast tacking into considerations a dynamic change of their conflicting degree We explored in this paper the problem of adjusting broadcast programs to cope with the variation of conflicting probability during transactions execution. By exploiting the features of the basement adjustment and the smooth adjustment, we proposed a heuristic based algorithm ADDUcast to adjust abundant data allocations therapy satisfying as much transactions as possible. Performance of algorithm ADDUcast was analyzed and a system simulator was developed to validate our results. It was shown by our simulation results that the allocation achieved by algorithm ADDUcast are highly maintain the freshness of data items with reduced response time. This feature and the efficiency of algorithm ADDUcast justify its practical importance.

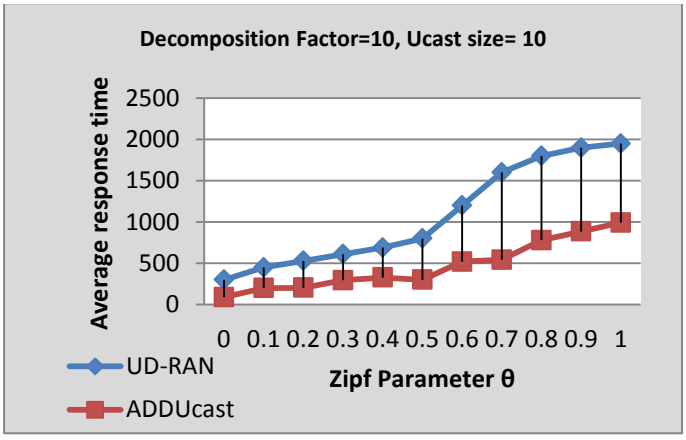


Fig. 7. Average response time with avergae Ucast size

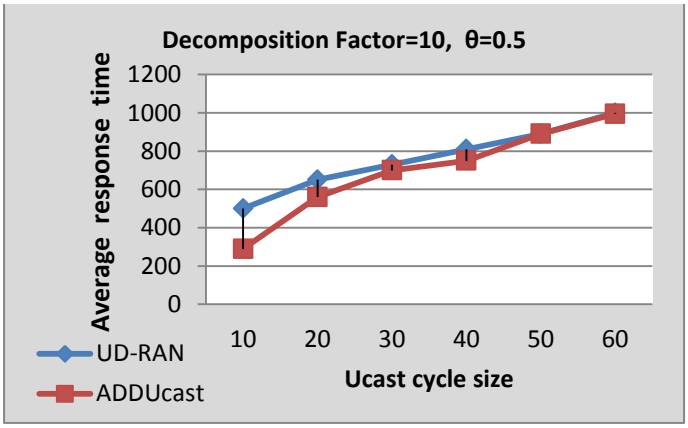


Fig. 8. Average response time under different Ucast cycle size

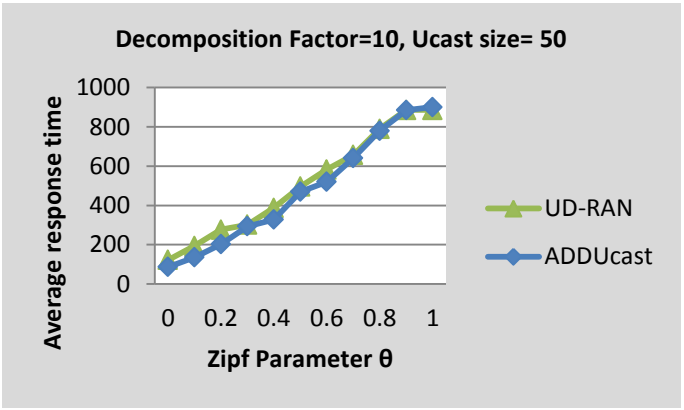


Fig. 9. Average response with large Ucast cycle

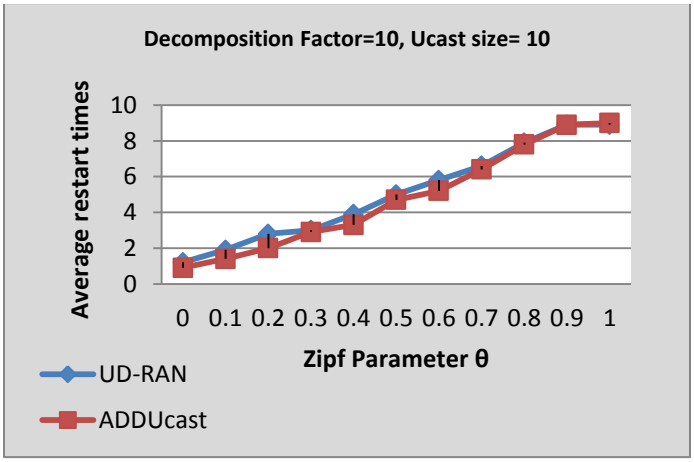


Fig. 10. Average restart times with avergae Ucast size

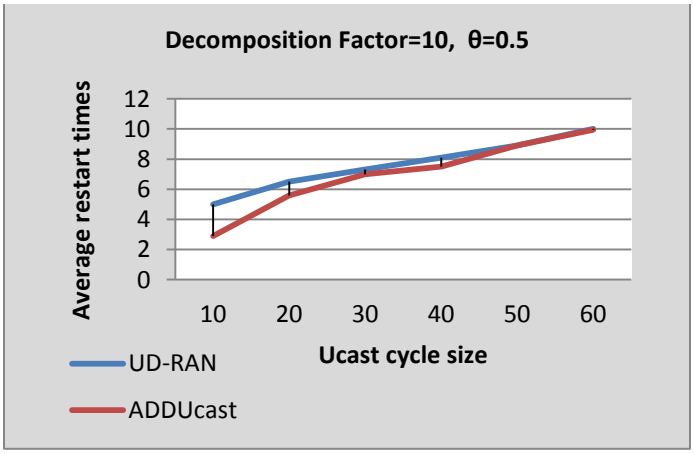


Fig. 11. Average restart times under different Ucast cycle size

ACKNOWLEDGMENT

This research is funded by the Deanship of Research and Graduate Studies in Zarqa Private University /Jordan"

REFERENCES

[1] Acharya, S., Alonso, R., Franklin, M., and Zdonik, S., "Broadcast disks: data management for asymmetric communication environments," Proc. of the ACM SIGMOD Conference, pp.199-210, 1995.

[2] Bernstein, P. A., Hadzilacos, V., and Goodman, N., Concurrency control and recovery in database systems, Addison-Wesley Publishing Company, 1987.

[3] Ozsu, M. T., and Valduriez, P., Principles of distributed database systems, Prentice Hall, 1991.

[4] Lam, K. Y., Au, M. W., and Chan, E., "Broadcast of consistent data to read-only transactions from mobile clients," Proc. of the Second IEEE Workshop on Mobile Computing Systems and Applications, 1999.

[5] Pitoura, E., "Supporting read-only transactions in wireless broadcasting," Proc. of the DEXA98 International Workshop on Mobility in Databases and Distributed Systems, pp. 428-422, 1998.

[6] Pitoura, E., and Chrysanthis, P. K., "Scalable processing of read-only transactions in broadcast push," Proc. Of the 19th IEEE International Conference on Distributed Computing System, 1999.

[7] Lee, SangKeun, Hwang, Chong-Sun, Kitsuregawa, Masaru,. Efficient, energy conserving transaction processing in wireless data broadcast. IEEE Transactions on Knowledge and Data Engineering 18 (9), 1225–1237. September 2006.

[8] Lee, V., Lam, K., Son, S.H., Chan, E, On transaction processing with partial validation and timestamps ordering in mobile broadcast environments. IEEETransactions on Computers 15 (10), 1196–1211 2002.

[9] Cho, H. Concurrency control for read-only client transactions in broadcast disks. IEICE Transactions on Communications E86-B (10), 3114–3122. 2003

[10] Lee, Victor C.S., Lam, Kwok Wa, Kuo, Tei-Wei, Efficient validation of mobile transactions in wireless environments. Journal of Systems and Software, 183–193. 2004.

[11] Lee, SangKeun, Hwang, Chong-Sun, Kitsuregawa, Masaru. Using predeclaration for efficient read-only transaction processing in wireless data broadcast. IEEE Transactions on Knowledge and Data Engineering 15 (6), 1579– 1583. Nov/Dec, 2003.

[12] Imielinski, T., and Viswanathan, S., and Badrimath, B., "Energy efficient indexing on air," Proc. of the ACM SIGMOD Conference, 1994.

[13] Huang, Y., and Lee, Y. H., "Concurrency control protocol for broadcastbased transaction processing and correctness proof," ISCTA PDCS 2001,in press, August 2001.

[14] W.-C. Peng and M.-S. Chen. Dynamic Generation of Data Broadcasting Programs for a Broadcast Disk Array in a Mobile Computing Environment. In Proceeding of the ACM 9th International Conference on Information and Knowledge Management, pages 38—45, November 2000.

[15] Jiun-Long Huang, Ming-Syan Chen: Dynamic Leveling: Adaptive Data Broadcasting in Mobile Computing Environment.MONET 8(4): 355-364 (2003)

[16] Sunggeun Park, Sungwon Jung; An energy-efficient mobile transaction processing method using random back-off in wireless broadcast environments The Journal of Systems and Software vol 82 pp 2012–2022, 2009

[17] Vikas Goel, Ajay Kumar Anil Kumar Ahlawat; A Comparative Study of Energy Efficient Air Indexing Techniques for Uniform Broadcasting International Journal of Computer Applications COMNET-2011

AUTHOR PROFILE

Ahmad Alqerem obtaining a BSc in 1997 from JUSTUniversity and a Masters in computer science from Jordan University in 2002. PhD in mobile computing at Loughborough University, UK in 2008. He is interested in concurrency control for mobile computing environments, particularly transaction processing. He has published several papers in various areas of computer science. After that he was appointed a head of internet technology Depts. Zarka University.

Exon_Intron Separation Using Amino Acids Groups Frenquency Repartition as Coding Technique

Afef Elloumi Oueslati
Unite Signal, Image et Reconnaissance de Formes,
Département de Génie Electrique, ENIT, BP 37, Campus Universitaire, Le Belvédère, 1002, Tunis

Noureddine Ellouze
Unite Signal, Image et Reconnaissance de Formes,
Département de Génie Electrique, ENIT, BP 37, Campus Universitaire, Le Belvédère, 1002, Tunis

***Abstract*—this paper presents a new coding technique based on amino acids repartition in chromosome. The signal generated with this coding technique constitutes, after treatment, a new way to separate between exons and introns in a gene. The algorithm proposed is composed of six steps. We convert from ATCG to amino acids. We specify the amino acid order group. We constitute the signal based on group's repartition. We apply a smoothing technique on resulting signal. We inverse the exons peaks from minima to maxima. We show the separation between exons and introns regions. We present here the results obtained on the gene reference G 56F11.**

Keywords—exons; introns; amino acid coding technique; amino acid repartition; exons - introns separation

I. INTRODUCTION

Genomic signal processing consists in applying signal processing method to code and analyze the genome. These analyses can be made on DNA sequences, RNA sequences or proteins. All of them are represented by characters. DNA and RNA are represented by four characters and proteins are made of twenty letters. The succession of the DNA's bases: A, G, C, and T, constitutes the hereditary message. Each DNA fragment involves a specific protein synthesis process. A set of 20 different amino acids synthesize proteins, following subsequent order of three bases called codon. A total of 64 different combinations specify 20 amino acids and three stop codons, namely TAA, TAG, and TGA. Signal processing methods have focused on genomic signals analysis and many methods of coding and analyzing are proposed [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. The application here concerns the protein coding regions (exons) and non-coding regions (introns). These regions are sometimes located by interpreting coding techniques results directly. But generally, the revealing periodicities are necessitating some transforms to make decision. In fact, the weighted real and complex value proposed by Anastassiou in [18, 19] needs an analyzing technique to decide on the sequences' category. This coding technique needs the spectral analysis to enhance the specific features of exonic regions [6, 7, 9, 10, 11, 20, 21, 22, 23, 24, 25, 26, 27]. The DNA walks consist in summing the values of each base or nucleotides' class [28, 29, 30]. A grammar of alphabets defines the location of each base in a codon or three bases [31]. The tetraedrical representation reveals the exon's characteristics and contributes to the identification of these regions [32, 33]. The aim of each coding method is to improve the hidden information for further analysis. In this context, statistic analysis has been elaborated to define the similitude degree and the region's dependence. The purpose is to find differences between the coding and non coding regions in the DNA sequence. Li demonstrates the existence of a correlation between all the regions [34, 35]. Peng et al. show on the other hand that the long range correlation is related to the intronic regions [36]. We focus here on periods based methods to analyze data. A method, called periodicity transforms proposed by Sethares and Staley detects periodicities in data with projections onto periodic subspaces [37]. This technique isn't based on a predefined set basis such as Wavelet or Fourier Transform. It associates different algorithms to calculate a specific set of non orthonormal basis elements which are directly depending on the analyzed data. This technique reveals periodicities relevant to signals by decomposition into the basic periodic components. A method based on amino acid coding for coding region detection followed by principal component analysis and wavelet transform is proposed in [38]. In our previous work presented in the reference [39], we focus on a particular and a specific periodicity, so we propose to apply the pitch synchronous analysis on genomic DNA sequences. This technique is based on the wavelet transform. The proposed method in this paper is dealing with a coding technique based on amino acids and its frequency repartition in the chromosome. In fact, the protein sequence is traduced into numerical signal by replacing each amino acid group by its frequency repartition. We demonstrate that such signal is an efficient tool to separate between exon and intron in a gene.

The paper is divided into 5 sections. Section 2 describes the coding technique, detailing the frequency order groups of amino acids used. Section 3 presents the methodology used to convert the signal obtained by the coding technique to a segmenting tool allowing the separation of exon and introns. Section four illustrates the accuracy of the proposed method by presenting the results obtained. The last section concludes the paper.

II. THE CODING TECHNIQUE BASED ON AMINO ACIDS GROUPS REPARTITION

The coding technique proposed is based on the probability (frequency) of apparition of each amino acid group in the entire chromosome. The repartition order corresponds to the number of amino acid in a group. The order one specify the probability of repartition of each amino acid (AA) as: A, D, M, N…. The order 2 is related to a combination of two consecutives AA as MN, MA, NY,…. The third order is for three successive AA such MNA, ADA,…..

These probabilities are calculated by the following equation

$$Paag = Naag / Naach \quad (1)$$

With Paag represents the Probability of one amino acid group, N*aag represents the number of one amino acid group in the entire chromosome and* N*ach represents the number of all amino acids in the chromosome*

For example:

Order 1:

For amino acid M $Pm = Nm / Naach;$

For amino acid Y $Py = Ny / Naach$

Order 2:

For amino acid group AM $Pam = Nam / Naach;$

For amino acid group YZ $Pyz = Nyz / Naach$

Order 3:

For amino acid group DTS $Pdts = Ndts / Naach$

For amino acid group YYY $Pyyy = Nyyy / Naach$

To code the DNA sequence, we use the probabilities calculated for each order. In fact, each value in the numeric signal is obtained by replacing each position k of one amino acid group by its probability of repartition values as expressed in equation 2

$$Saa(k) = \Sigma_i \; Paag(i,k) \quad (2)$$

The indice i determines the aag group and k represent the position in the sequence to code

The entire numeric signal is then obtained by calculating the sum of these probabilities on the entire sequence as expressed in equation 3

$$Sa = \Sigma_\kappa \; Saa(k) \quad (3)$$

The probabilities are calculated on the entire chromosome but the numeric signal can be applied only on the considered sequence to code, in our case the sequence is the gene. When the sequence is coded, the next step consists in dealing with the numeric signal to show the separation between exons and introns in a gene.

III. The Exon Introns Separation Methodologie

The aim of the proposed coding technique is to distinguish between exons and introns in a gene. Using the probability of repartition into the coding technique is a way to avoid the analysis method. In fact, the numerical signal obtained after the coding is able to separate between the considered regions. We just add some preprocessing techniques. The methodology proposed begins from chromosome with its ATCG form and finishes with highlighting the exons-introns limits in a gene. The approach proposed is divided into six parts as follows:

- Converting chromosome from the ATCG form to amino acid form via the genetic code ATATCGATCTG→ISI*

  With *represents stop codons

- Specifying the repartition order which corresponds to the number of amino acid used in a group. Order 1 for one amino acid, order 2 for two amino acids…etc.

- Calculating the probabilities Paag on the entire chromosome. We present in the Table1 the Paag for each amino acid for the order 1. Table 2 presents some examples of the Paag for the order two.

TABLE I. Amino Acid Probabilty For Order 1

Amino acid group (order 1)	*Paag*	*Amino acid group(order 1)*	*Paag*
L	0.0940	Q	0.0349
F	0.0929	P	0.0349
S	0.0866	G	0.0348
K	0.0830	A	0.0342
I	0.0724	Y	0.0292
R	0.0601	C	0.0272
N	0.0553	H	0.0252
T	0.0461	D	0.0238
V	0.0460	M	0.0149
E	0.0398	W	0.0126
*	0.0520	-	-

TABLE II. Examples of Amino Acid Probabilties for Order 2

Amino acid group (order2)	*Paag*
FF	0,0115
FL	0,0107
FS	0,0103
KK	0,0102
IF	0,0099
KI	0,0095
LL	0,0094
LK	0,0089
SS	0,0085
KL	0,0079
NF	0,0078

- Coding the gene with the probabilities of the specified order. For example the numeric sequence of seqaa=ISI* for the first order is

  seqn= 0.0724 0.0866 0.0724 0.0520

- The resulting signal needs smoothing to reveal clearly the different region in a gene. We choose simply to apply the mean value on a number of consecutive neighbors. In our case we fix the number to 18.

  Each value k is replaced by the mean value of its 18 neighbors as expressed in equation 4

  $$Smean(\kappa) = \Sigma i\ Saa\ (i) \qquad (4)$$

  With index i is varying from k to k+18. The smoothed signal is the sum for all the values k, the expression is given by equation 5

  $$Ssmooth = \Sigma\kappa\ Smean\ (k) \qquad (5)$$

  With the index k is varying from 1 to the length of the gene.

  Finally, the obtained signal shows that exons are characterized with minima so we inverse signal to have the exons as maxima with the equation 6:

  $$Sfinal = 1 - Ssmooth \qquad (6)$$

- The Sfinal of equation 6 is the signal on which we distinguish the exons and introns. In fact, the peaks represent the exons regions and the separation is clear.

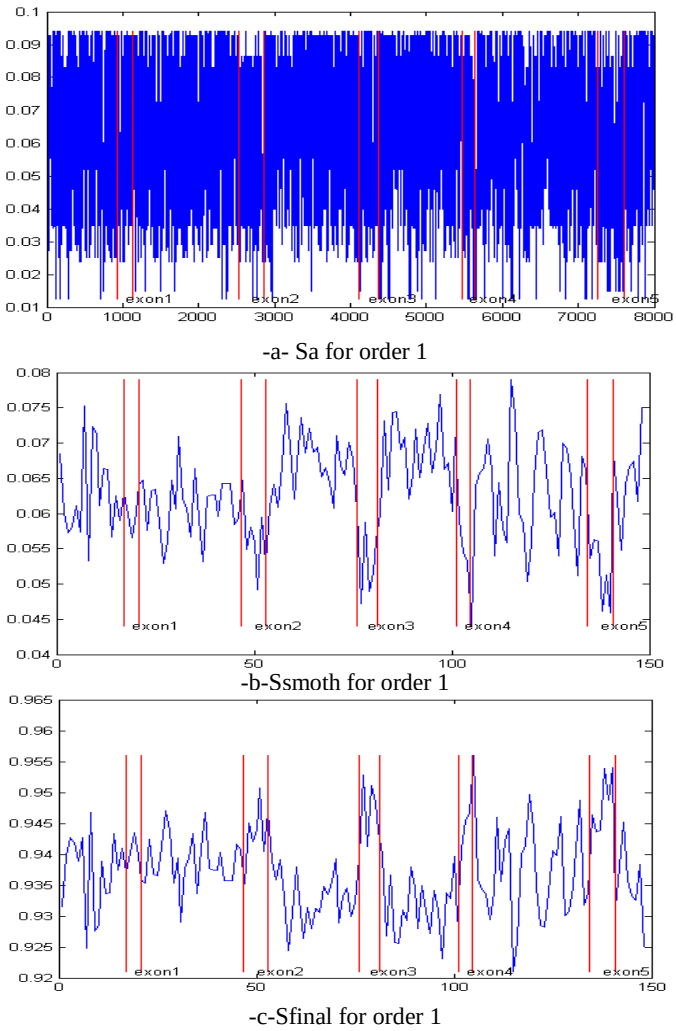
We test in our analysis different orders. We illustrate here the results for orders from order 1 to order 4. We remind that the order 4 is a combination of four consecutive amino acids. We present here the results obtained for these 4 orders for the reference gene G56F11. In the table III, we present the exon's positions for the five exons initially and after applying the mean value as smoothing technique with the mean value obtained for 18 neighbors.

TABLE III. EXONS' POSITION FOR GENE G56F11

Exon's number	Exon's position	
	Initial	*After mean value*
1	929-1135	17-21
2	2528-2857	46-52
3	4114-4377	76-81
4	5465-5644	101-104
5	7255-7605	134-140

The methodology's steps are illustrated by the figure 1for the reference gene G56F11.

For each order from order 1 to order 4, we consider 3 figures. The methodology's steps are illustrated by figure 1.



-a- Sa for order 1

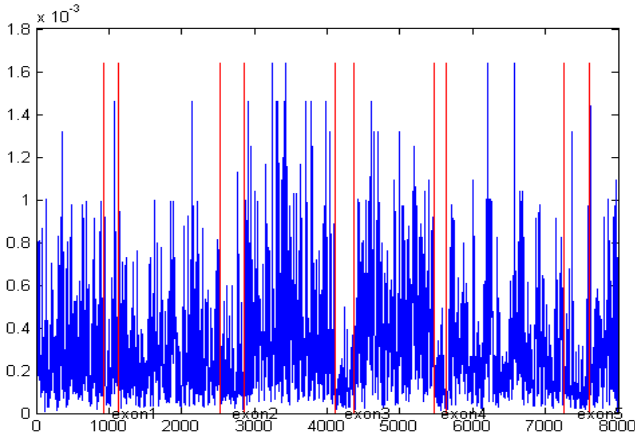
-b-Ssmoth for order 1

-c-Sfinal for order 1

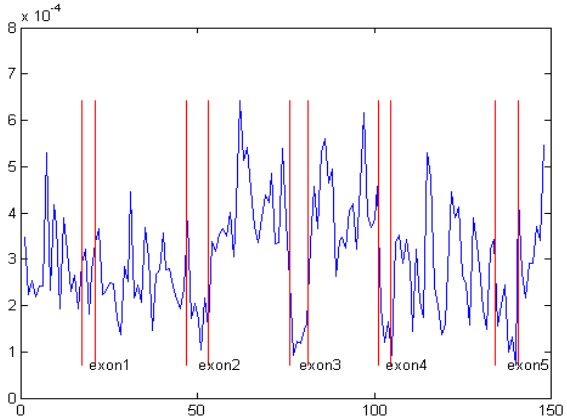
Fig. 1. The methodology's steps for order 1. –a- is the signal Sa, -b- represents Ssmooth and the –c- is the Sfinal signal

We present in each figure the results obtained for each order. Figure 2 is related to order 2. Figure 3 exposes order 3 results and figure 4 the exon intron separation with the probabilities repartition of order 4.

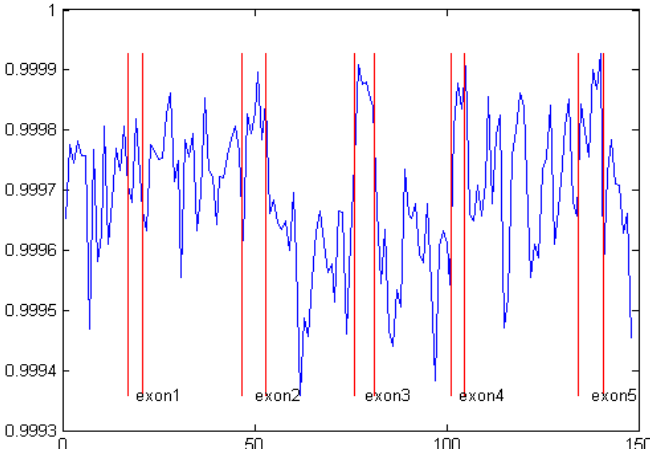
These figures are subdivided as follows. The first one is the signal corresponding to the probabilities repartition Saa. It is calculated for the whole chromosome and applied to the gene G56F11. The initial exon's positions are delimited with red lines. The second signal is the smoothed one Ssmooth. The modified exon's positions after the mean value, given in table 3, are represented with red lines to highlight the regions separation. The third subfigure is the final signal enhancing the exon's peaks as maxima to clearly present the exons _introns separation.



-a- Sa for order 3

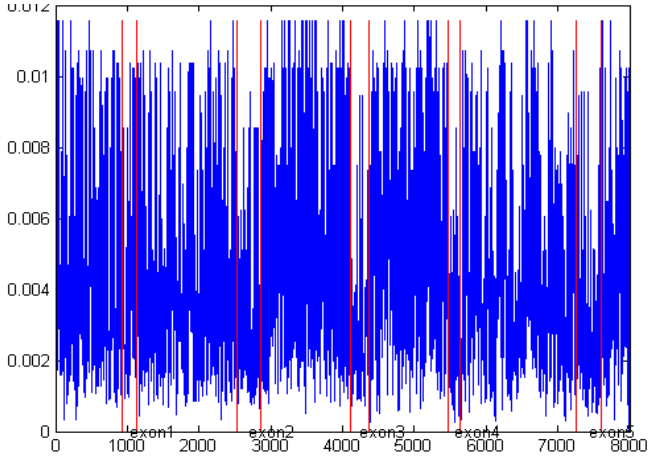


-b- Ssmoth for order 3

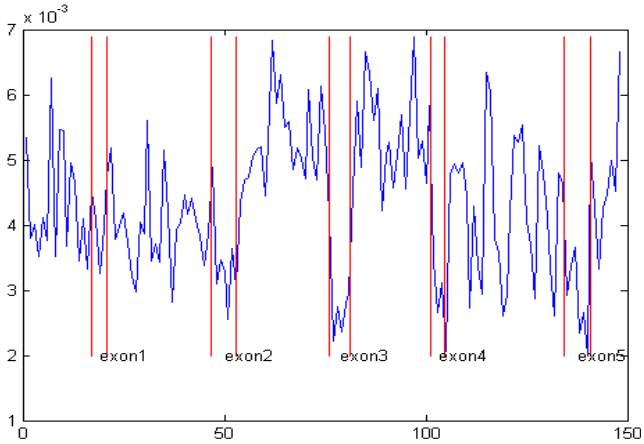


-c- Sfinal for order 3

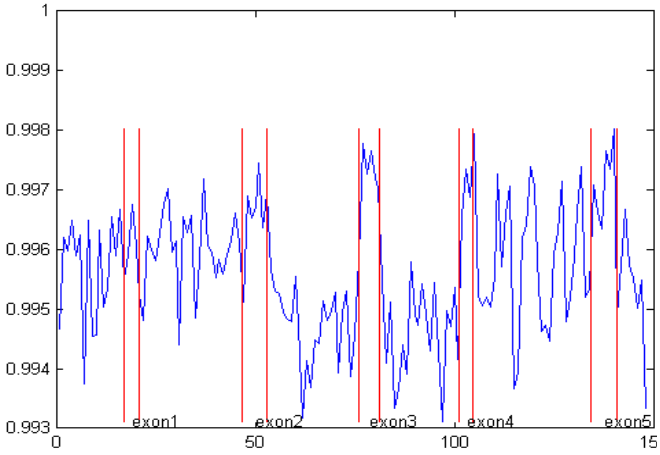
Fig. 2. The methodology's results for order 2. –a- is the signal Sa, -b- represents Ssmooth and the –c- is the Sfinal signal



-a- Sa for order 2

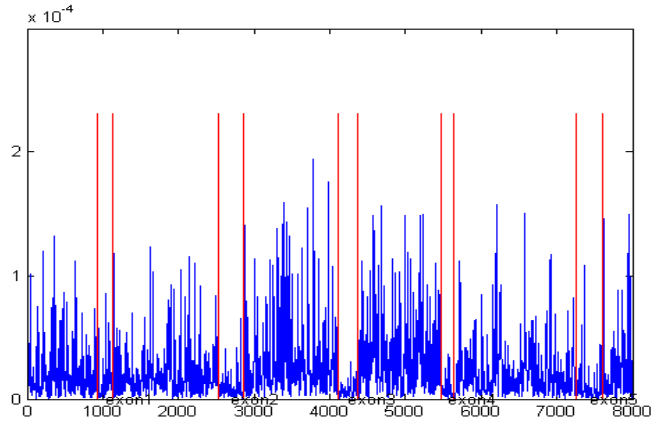


-b- Ssmoth for order 2

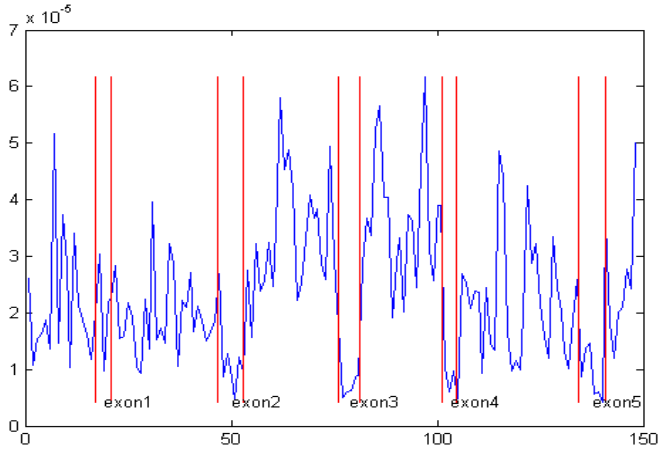


-c- Sfinal for order 2

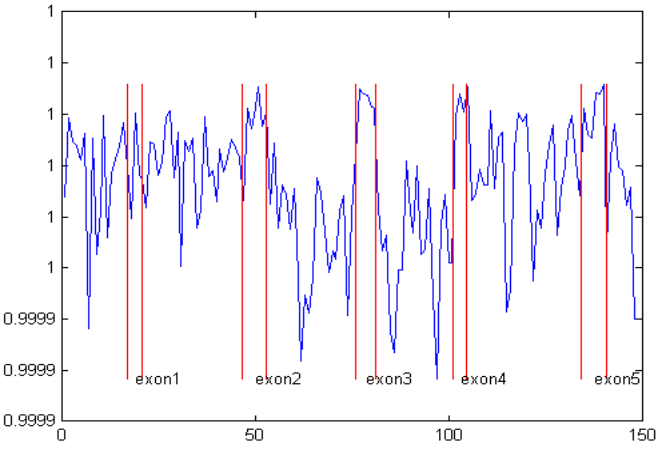
Fig. 3. The methodology's results for order 3. –a- is the signal Sa, -b- represents Ssmooth and the –c- is the Sfinal signal



-a- Sa for order 4



-b- Ssmoth for order 4



-c- Sfinal for order 4

Fig. 4. The methodology's results for order 4. –a- is the signal Sa, -b- represents Ssmooth and the –c- is the Sfinal signal

REFERENCES

[1] C. Mathé, M.F. Sagot, T. Schiex, and P. Rouzé, "Current methods of gene prediction, their strengths and weaknesses," Nucleic Acids Research, 2002, vol. 30, no. 19, pp. 4103–4117,

[2] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," Nucleic Acids Research, 1982, vol. 10, no. 17, pp. 5303–5318.

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, UK, 1998.

[4] N. Chakravarthy, A. Spanias, L. D. Iasemidis, and K. Tsakalis, "Autoregressive modeling and feature analysis of DNA sequences," EURASIP Journal on Applied Signal Processing ,2004, vol. 2004, no. 1, pp. 13–28,.

[5] J. V. Lorenzo-Ginori, A. Rodrıguez-Fuentes, R. G.Abalo, and R. S. Rodrıguez, "Digital signal processing in the analysis of genomic sequences, "Current Bioinformatics, 2009, vol.4,no.1,pp.28–40.

[6] S. Nancy Yu and Y. Hong. Short exon detection in DNA sequences based on multifeature spectral analysis. EURASIP Journal on Advances in Signal Processing, 2011.vol no pp

[7] 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 , 1997, vol. 13, no. 3, pp. 263–270.

[8] M. Akhtar, E. Ambikairajah, and J. Epps, "Optimizing period- 3 methods for eukaryotic gene prediction," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08), 2008, pp. 621–624.

[9] H. Yan and T. D. Pham, "Spectral estimation techniques for DNA sequence and microarray data analysis, 2007, "Current Bioinformatics, vol. 2, no. 2, pp. 145–156,.

[10] M. K. Choong, and H. Yan, Multi-scale parametric spectral analysis for exon detection in DNA sequences based on forward-backward linear prediction and singular value decomposition of the double-base curves. Bioinformation, 2008. 2(7), 273-278.

[11] R. Jiang and H. Yan, "Studies of spectral properties of short genes using the wavelet subspace Hilbert-Huang transform (WSHHT)," Physica A , 2008, vol. 387, no. 16-17, pp. 4223–4247.

[12] T. P. George and T. Thomas, "Discrete wavelet transform denoising in eukaryotic gene splicing," BMC Bioinformatics, 2010, vol.11, supplement 1, article S50.

[13] Y.Wu, A.W.-C.Liew,H.Yan,andM.Yang,"DB-Curve: a novel 2D method of DNA sequence visualization and representation," Chemical Physics Letters , vol. 367, no. 1-2, pp. 170–176, 2003

[14] M. Akhtar, J. Epps, J. and E. Ambikairajah, E. On DNA numerical representations for period-3 based exon prediction. In Genomic Signal Processing and Statistics, 2007. GENSIPS 2007. IEEE International Workshop on (pp. 1-4). IEEE.

[15] H.T chang , C.J Kuo, N.W . Lo and W.Z Lv. DNA sequence Representation and comparison Based on Quaternion Number System. International Journal of Advanced Computer Science & Applications. 2012 vol 3, no 11

[16] K.S. Sathish and N. Duraipandian. An effectice identifcation of Species from DNA Sequence: A Classification Technique by integrating DM and ANN. International Journal of Advanced Computer Science & Applications. 2012 vol 3, no 8

[17] S.N. Devi and S.P. Rajagopalan. A study on Feature Selection Techniques in Bio-Informatics. International Journal of Advanced Computer Science & Applications. 2011 vol 2, no 1,pp 138-144

[18] D. Anastassiou. "Genomic signal processing". Signal Processing Magazine, IEEE, 2001, vol. 18, no 4, p. 8-20.

[19] D. Anastassiou. "DSP in genomics: processing and frequency-domain analysis of character strings". In : Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on. IEEE, 2001. p. 1053-1056.

[20] R. P Costa," Gene prediction algorithms". Computational Biology, 2003, p. 1-7.

[21] A. Elloumi, Z. Lachiri and N. Ellouze. « DNA sequence analysis: From DNA sequencing to gene prediction". In : 17th IMACS World Congress Scientific Computation, Applied Mathematics and Simulation. 2005.

[22] A. Elloumi, Z. Lachiri and N. Ellouze. Spectral Analysis of DNA Sequence: The Exon's Location Method. In : Digital Signal Processing, 2007 15th International Conference on. IEEE, 2007. p. 115-118.

[23] A. Elloumi, Z. Lachiri and N. Ellouze. 3D Spectrum Analysis of DNA Sequence: Application to Caenorhabditis elegans Genome. In :

Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on. IEEE, 2007. p. 864-871.

[24] Y. Changchuan and Y. S.-T Stephen. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. Journal of theoretical biology, 2007, vol. 247, no 4, p. 687-694.

[25] J. Xianyang, D. Lavenier and Y. S.-T Stephen. Coding region prediction based on a universal DNA sequence representation method. Journal of Computational Biology, 2008, vol. 15, no 10, p. 1237-1256.

[26] A. S Marhon, and S. C. kremer. Gene prediction based on DNA spectral analysis: a literature review. Journal of Computational Biology, 2011, vol. 18, no 4, p. 639-676.

[27] Y. Changchuan and Y. S.-T Stephen. A Fourier characteristic of coding sequences: origins and a non-Fourier approximation. Journal of Computational Biology, 2005, vol. 12, no 9, p. 1153-1165.

[28] J. A. Berger, S. K. Mitra, M. Carli, Marco, and al. New approaches to genome sequence analysis based on digital signal processing. University of California, 2002. Workshop on Genomic Signal Processing and Statistics (GENSIPS), IEEE, Raleigh, North Carolina, October,pp. 1–4, .

[29] J. A. Berger, S. K Mitra, M. Carli, Marco, and al. Visualization and analysis of DNA sequences using DNA walks. Journal of the Franklin Institute, 2004, vol. 341, no 1, p. 37-53.

[30] J. A. Berger, S. K. Mitra and M, J. Astola. Power spectrum analysis for DNA sequences. In : Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on. IEEE, 2003. p. 29-32.

[31] D. Nicorici and J. Astola. Segmentation of DNA into coding and noncoding regions based on recursive entropic segmentation and stop-codon statistics. EURASIP Journal on Applied Signal Processing, 2004, vol. 2004, p. 81-91.

[32] P.D. Cristea. Large scale features in DNA genomic signals. Signal Processing, 2003, vol. 83, no 4, p. 871-888.

[33] P.D. Cristea. Multiresolution phase analysis of genomic signals. In : Control, Communications and Signal Processing, 2004. First International Symposium on. IEEE, 2004. p. 743-746.

[34] W. Li and K. Kaneko. Long-range correlation and partial 1/fα spectrum in a noncoding DNA sequence. EPL (Europhysics Letters), 1992, vol. 17, no 7, p. 655.

[35] W. Li. The study of correlation structures of DNA sequences: a critical review. Computers & chemistry, 1997, vol. 21, no 4, p. 257-271.

[36] C.K. Peng, S.V. Buldyrev,A. L. Goldberger and al. Long-range correlations in nucleotide sequences. Nature, 1992, vol. 356, no 6365, p. 168-170.

[37] W.A. Sethares and W. Thomas. Periodicity transforms. Signal Processing, IEEE Transactions on, 1999, vol. 47, no 11, p. 2953-2964.

[38] C-Y. Tsai and C-C Chiu. An efficient conserved region detection method for multiple protein sequences using principal component analysis and wavelet transform. Pattern Recognition Letters, 2008, vol. 29, no 5, p. 616-628.

[39] A. Elloumi, Z. Lachiri and N. Ellouze. Detecting particular features in C. elegans genomes using Synchronous Analysis based on Wavelet Transform. International Journal of Bioinformatics Research and Applications, 2011, vol. 7, no 2, p. 183-201.

Methods of Isolation for Application Traces Using Virtual Machines and Shadow Copies

George Pecherle
Faculty of Electrical Engineering and Information Technology, University of Oradea Oradea, Romania

Cornelia Győrödi
Faculty of Electrical Engineering and Information Technology, University of Oradea Oradea, Romania

Robert Győrödi
Faculty of Electrical Engineering and Information Technology, University of Oradea Oradea, Romania

***Abstract*—To improve the user's experience, almost all applications save usage data: web browsers save history and cookies, chat programs save message archives and so on. However, this data can be confidential and may compromise the user's privacy. There are third party solutions to automatically detect and wipe these traces, but they have two problems: they need a constantly updated database of files to target, and they wipe the data after it has been written to the disk. Our proposed solution does not need a database and it automatically reverts the application to its initial (clean) state, leaving no traces behind. This is done by using a monitoring process developed by us and the Volume Shadow Copy Service that takes snapshots when the application runs and restores them at the end of the run.**

***Keywords*—security; privacy; application traces; data wiping; virtual machines; shadow copies; sandbox**

I. Introduction

Storage capacity of disks has increased during the recent years - sometimes exponentially - facilitating a large number of programs to work together and make sometimes very complex operations, and also facilitating the amount of data that these programs work with. Therefore, for the user of a modern computer system, it has become impossible to know or manually check the data and the software stored on a computer system, for reasons that relate to the huge volume that is stored and to the way programs hide and/or encrypt data during their normal operations. It is noted in this context, that there is a strong need to protect the data stored on a computer system against external agents that might compromise the security without the user's knowledge, to ensure the user's privacy and a proper functioning of the operating system.

This protection was achieved by designing modern operating systems and even computer systems to avoid vulnerabilities to external factors and facilitate the implementation of subsystems designed for maintaining the security of the data.

The study from [1] shows that any system has vulnerabilities, 14,900 files being detected as files that are part of programs designed to attack an Android operating system, based on the Linux kernel, that has not traditionally been the target of attacks until recently.

Another important issue that comes up is that programs can leave traces of their usage that contain private information about the user. This is why most Internet browsers have features that allow the user to start private sessions that don't save usage traces, such as data about the accessed sites, passwords and other data entered in web forms [2] or so-called cookies used for saving sessions on specific portals that require registration or other information to identify the user, features that improve the user's experience on the web. The need to solve this problem started to become real, especially because of websites that lead the user to expose more data to the Internet browser (such as online shopping sites, flight bookings, banking services, etc.), very often on devices that do not belong to the user and that can be accessed by other people, programs or sites that are not trustable.

II. Current Solutions To Problems of Data Security and Private Data Protection

There are a lot of software methods to ensure data protection (both in terms of system reliability and data security) and they are implemented in various combinations by the operating systems and by specialized software.

Along with the methods and mechanisms that ensure data security and fault tolerance of the system (built into the operating system), there are also methods of protection provided by third parties. These programs are either actively working to detect problems generated in the system (such as suites of programs that detect viruses and other malware, or software that verify the integrity of data or of different subsystems), or programs that are passive. The experience shows that none of these methods can fully respond to security and privacy needs of the user. What differentiates security solutions from the user's perspective is their ability to either prevent a problem, or try to solve it after it has been generated.

Of course, users prefer the first method (preventive security) and this paper is about this type of security protection. Two methods will be presented and then how we used them to design our privacy protection system, that isolates private data using virtual machines and shadow copies.

A. Virtual Machines

A good compromise between performance / data accessibility and data security can be made by using virtual machines that are essentially computer systems with a similar behavior as the original (physical) ones. They are actually abstract implementations of real systems [3], and because of this reason, they can be protected against external factors, by filtering the communication ways with the outside environment, and also by the fact that the state of these machines can be saved periodically so there is always the

possibility of rolling back to a pre-infection state. Virtual machines can benefit from the existence of more processor modes, the supervisor mode of the virtual machine being named the "hypervisor".

This method, of using virtual machines, although it is useful in some situations - for example when you want to obtain a controlled environment to implement solutions with specific purposes, of "closed box" type [4] - has only minor advantages from a security point of view, compared to an operating system on a real machine. This is because, in most scenarios, the virtual machine cannot be completely isolated from the outside environment in order to meet the usage guidelines. And the more the accessibility of virtual machines increases, the more the security level of virtual machines becomes closer to that of a real system.

B. Sandboxes

A good method to secure the data and the operating system, that is closer to using virtual machines, are the so-called "sandboxes". These are programs that allow the execution of other programs (called host programs) with a limited and controlled set of resources.

These sandboxes work in different ways, depending on the purpose they were designed for. A recent trend that is worth mentioning is that part of the modern and the most successful operating systems, low-level sandboxes (close to the operating system layer) have been implemented for a large amount of applications [5].

Using a sandbox for a secure system can be achieved very easy, by installing and using a virtual machine with an operating system installed, for example Oracle VM Virtual Box [21].

III. Comparison With Other Similar Methods

There are a lot of solutions (especially software programs) that can automatically detect and securely delete traces of application usage. There are usually two types of solutions:

- Solutions that detect application traces and react to this phenomenon AFTER the data has been written to disk (securely erase it), sometimes long after that (e.g. when the user launches the erase process), leading to serious privacy concerns.
- Solutions that detect application traces and react to this phenomenon BEFORE the data is written to the disk. This is the most efficient and secure method because there is a very low risk for sensitive data to fall into the wrong hands.

Our method described in this paper falls into the second category and its originality comes from the way modified data is intercepted and handled: saving snapshots of the original data using the Volume Shadow Copy Service then restore it at the end. Before restoring the original data, a secure erase of the modified files could be implemented (we could make a list of modified files, using the preparation callback routines of the minifilter driver).

We have also identified another method that is also from the second category above and it is called Sandboxie [19]. This solution saves application traces in a special sandbox, not on their original locations. This way, sensitive data can be erased all at once, from the same place.

A problem we have detected with our method is: what if the user wants to save data in a file and that should remain modified? Taking into consideration the idea from the Sandboxie solution described above, we could modify our solution to declare a "safe" area, where this data can be saved and that should be the user's responsibility to clean after he no longer needs that data.

A limitation of our solution, but also of many other available solutions is that it's only for Windows operating systems.

The disadvantages of the solutions from the first category (just detect where applications save data and securely erase it) are obvious: there will always be a delay between saving sensitive data and erasing it. During this time, the data can fall into the wrong hands. Also, the locations where applications save data change rapidly, so a constantly updated database of locations needs to be maintained. And this can lead to sensitive data not being caught and erased.

Also, another method that is worth mentioning and that was previously proposed by me, implements an algorithm that determines the sensitivity of files using a pre-defined set of rules made by the user. These rules are self-adaptable, in a way that they can improve themselves, taking into account patterns detected in other files securely erased by the user [20].

IV. Proposed Sandboxed Solution To Isolate Private Data

The research we have performed and that will be presented here is a sandbox whose purpose is to isolate only files that are accessed by the host programs, using features already implemented in the Windows operating system. Some of them are features related to the file system, driver development for the Windows operating system and the Volume ShadowCopy Service (VSS).

The main benefit introduced by this research is the change of the rules imposed by a traditional sandbox, through controlled accessibility to files stored on the disk. The effects of this solution and also its efficiency in solving the data privacy problems will be presented next in this paper.

A. Motivation

In this section, we will present a research that demonstrates the flexibility and the simple way in which the security and protection methods of the operating systems can be extended. This design and implements a sandbox that is limited to protecting and isolating changes made by applications in the file system.

As shown previously, the full virtualization and very restrictive sandboxes don't always meet the security and data protection requirements of the users.

Moreover, the increased complexity of a virtual machine or that of a sandbox that abstracts levels very close to the kernel generates security problems that are similar to those of the operating system itself. Also, modern operating systems offer advanced protection mechanisms and policies that have been thoroughly tested. For this reason, it is enough to use features that are present in the operating system.

B. Requirements

Our research had to meet the following requirements:

- To be fully compatible with all recent Windows operating systems
- To run with minimal hardware resources
- To integrate with the operating system (through context menus, for example)
- To integrate with the graphical user interface of the host operating system
- To allow the user to choose the application to start in protected mode
- To prevent altering system files by the application that has started in protected mode
- To use a minimum amount of memory and to add a minimum wait time when the application is launched
- To avoid writing on the disk as much as possible
- To use a minimum amount of source code to avoid errors
- To use a minimum amount of source code that runs in supervisor mode
- The application should work transparently to the user

C. Specifications

The research we have done is based on a system application, and for this reason, we have used the C++ programming language that offers system oriented development features and that is very well supported by some of the Windows API functions, by their functionality and the documentation that is available for them. In order to develop a driver that we will use to monitor and block I/O requests with the file system, we have identified the filter drivers of the file system as being the type of drivers that we should use. These drivers had to be implemented in the C programming language. Another technology we have identified and that we have used to implement the file manipulation system is the VSS (Volume Shadow Copy Service). This technology was introduced in Windows Server 2003, it works with the NTFS file system and facilitates the creation of snapshots (saving the state of the file system on an NTFS volume), without affecting the normal operation of the system, using a technique similar to Copy-On-Write at block level [6] [7]. This technique operates at block level (not file level) and makes differential copies instead of full copies of the original data. First, a copy of the original data is created. Then, whenever a change to the original volume occurs, before this is written to disk, the block that is about to be modified is written to an area that stores "differences" to the original data. Using the original data blocks and the differences blocks, a shadow copy can be built that represents the copy of the data at the time it was created. The main advantage of this method is the speed, because it only writes the differences [18].

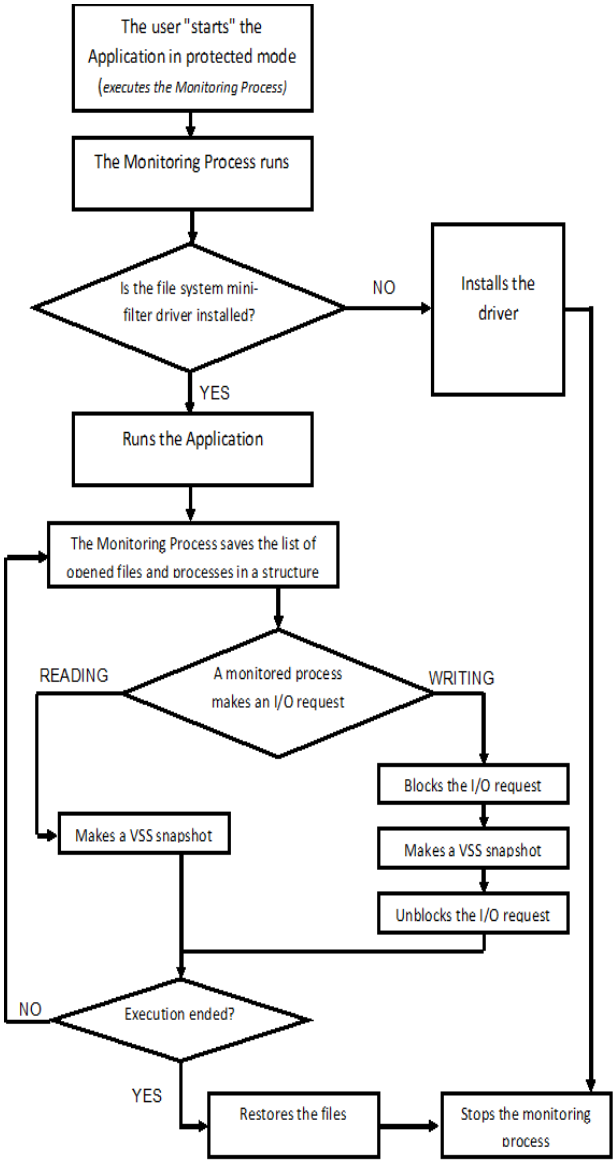


Fig. 1. The workflow chart of the monitoring process and its interaction with the protected (targeted) application

The minifilter driver can be implemented using a file system minifilter [8]. This makes it possible to write a small amount of source code for the supervisor mode, minimizing the risks of programming errors.

The concept of minifilter drivers is simple: a user makes a request for file I/O. Then the I/O manager sends the request to the file system. At that point, the filter manager intercepts the I/O request and calls the registered minifilters in their altitude orders. The altitude of a minifilter is a unique identifier that determines the order of attachment. The altitudes are allocated and managed by the operating system and it ensures that an instance of a minifilter driver is loaded at a location that is

appropriate to other instances of minifilter drivers. Also, minifilter drivers can register a preoperation callback routine, or a postoperation callback routine or both of them. Preoperation callback routines are called in descending order of altitudes (in case there are more minifilters drivers installed), then the I/O operation takes place, then the postoperation callback routines are called in ascending order of altitudes, from lowest to highest. [16]

To edit the C and C++ source code, we have used the integrated development environment of Microsoft Visual Studio 2012. In order to test and debug the program, we have used a virtual machine provided by Microsoft for Windows 7 - called the Windows XP Mode. The use of a virtual machine is just a method of protection of the system we have developed, on which we do not want to run untested code in supervisor mode.

Figure 1 describes the general workflow chart of the monitoring process and its interaction with the protected (targeted) application. One important thing to note is that the monitoring process will be launched when the user starts the protected (targeted) application.

If the file system minifilter driver is not present, the application will ask the user to install the driver before exiting. If the driver is present, the targeted application will be launched automatically. Then, a data structure will be created inside the monitoring process, to store a list of accessed files for each process of the targeted application. The monitoring process will continue to execute in parallel with the process (or the processes) of the targeted application. When an I/O request from one of the monitored processes is detected, the type of I/O request will be verified. For input operations, a preventive snapshot will be done (as optimization method). However, for output operations, the operation will be blocked until the Volume Shadow Copy Service performs the snapshot. For both situations (input and output operations), an internal verification will be done to detect when the execution of the targeted applications ends. If it's not ended, the monitoring process will continue to monitor the targeted application. If the targeted application ends, the original files will be restored using the Volume ShadowCopy Service snapshot and the execution of the monitoring process will end.

D. Implementation

The implementation consists of three separate components, each representing a separate project in Visual Studio:

- The monitoring process - that will monitor all I/O requests from the targeted application
- A service that will use SCM (Service Control Manager) [9] to start the monitoring process automatically and also to avoid User Account Control (UAC) messages, that elevate permissions for users. We will use the functions to install and uninstall the service.
- The file system minifilter driver

The documentation required for each of these components was obtained from:

- The documentation about writing Windows applications available from the MSDN (Microsoft Developer Network) website [10],
- The documentation about writing services in C++ using the Visual Studio 2012 development environment [11] and
- The documentation to write a file system minifilter driver, that is also available from MSDN [12]

The application has been integrated in the Windows graphical user interface using the Windows Registry, by extending a shell component (the main component of the Windows GUI), more exactly by changing a registry key from the Windows Registry as shown at [13]. The extension was done on the "exefile" subclass, and this means that a new context menu option will be present only for executable files and their shortcuts. We have called this context menu option, "Run in Protected Mode", as shown in Figure 2.

E. The Protected Mode Service

To start the monitored process in protected mode, we have used a service, that we called the Protected Mode Service. This is actually run in command line and can accept various command line parameters. The usage of the Protected Mode service is below:

```
pmservice [mode] [servicecommandstype] [command]
```

The [mode] parameters can accept 2 values: "service" or "process". If it's "process", normal process work is done. If it's "service", the next parameter is verified that it has one the following values:

- "config": the user can run some configuration commands, such as "query" (to retrieve and display the current service configuration, using the DoQuerySvc() function), "describe" (to update the service description to a default value, using the DoUpdateSvcDesc() function), "disable" (to disable the service, using the DoDisableSvc() function), "enable" (to enable the service, using the DoEnableSvc() function), "delete" (to delete the service, using the DoDeleteSvc() function).
- "control": the user can run some control commands, such as "start" (to start the service, if possible, using the DoStartSvc() function), "dacl" (to update the service DACL [14] to grant start, stop, delete, and read control access to the Guest account, using the DoUpdateSvcDacl() function), "stop" (to stop the service, using the DoStopSvc() function).
- "install": the service is installed in the Service Control Manager (SCM) database [15] by using the SvcInstall() function that we have implemented. Then a new entry ("Run in Protected Mode") is added in the context menu of executable files, by creating a special registry key under HKEY_CLASSES_ROOT, as below:

```
LPWSTR pm_command =
(LPWSTR)"C:\\dev\\ProtectedMode\\PMService.exe
process \"%1\"\0";

HKEY hkey;
DWORD  dwDisposition;
DWORD dwType, dwSize;

if (RegCreateKeyEx(HKEY_CLASSES_ROOT,
TEXT("exefile\\shell\\Run in Protected
Mode\\command"), 0, NULL, 0, 0, NULL, &hkey,
&dwDisposition) == ERROR_SUCCESS)
{

dwType = REG_SZ;
       dwSize = (wcslen(pm_command) + 1) *
sizeof(WCHAR);
       RegSetValueEx(hkey, NULL, 0, dwType,
(LPBYTE)&pm_command, dwSize);
       RegCloseKey(hkey);
}
```

And of course, we have implemented a class called ProcessMonitor, that has a constructor with the following arguments: ProcessMonitor(TCHAR *handlePath, TCHAR *monitoredProcPath) and that starts the monitored process as a child process, using the CreateProcess() function [17] and then makes the processing as described.

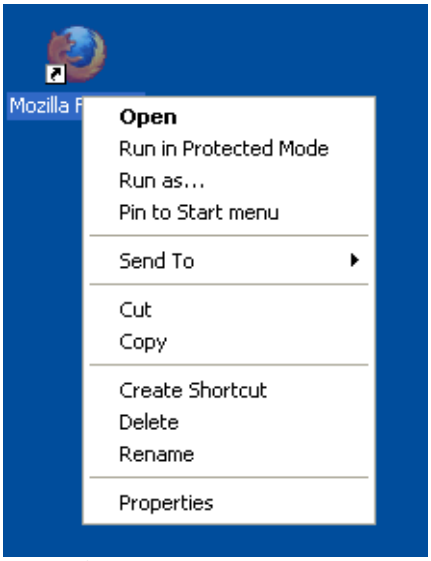


Fig. 2. Screenshot of the context menu with our new option "Run in Protected Mode" that will launch the monitoring process on this application

V. Conclusions

This paper implements a new way to protect the data manipulated by applications: isolating the data in a protected environment. This is a requirement because most applications have an uncontrollable and unpredictable way of saving their data and this can lead to privacy issues.

This is done using a system feature used for backup and system restore purposes, the Volume ShadowCopy Service, by doing a snapshot when the application makes the first I/O request and restoring it when the application ends.

The Windows operating system offers developers a set of programming interfaces that allow the extension of the system capabilities, also in supervisor mode, without the need to write long and complicated programs that are likely to have errors. The driver system, especially the filter drivers system, allows the extension of the functionality set for developers who need to obtain a different system behavior, by taking advantage of the hardware capabilities. On another note, the system offers security and protection measures that are meant to increase the system's reliability and the user's experience.

In the future, we would like to extend this research to mobile applications taking into account that data privacy on mobile devices is now a requirement of both home and corporate users.

References

[1] Y. Namestnikov, IT Threat Evolution: Q2 2012, Kaspersky Lab ZAO, http://www.securelist.com/en/analysis/204792231/IT_Threat_Evolution_Q1_2012.

[2] Private Browsing - Browse the web without saving information about the sites you visit, https://support.mozilla.org/en-US/kb/private-browsing-browse-web-without-saving-info.

[3] J. E. Smith and R. Nair, The architecture of virtual machines,Computer, vol. 38, nr. 5, pag. 32-38, 2005.

[4] T. Garfinkel, B. Pfaff, J. Chow, M. Rosenblum and D. Boneh, Terra: A virtual machine-based platform for trusted computing, ACM SIGOPS Operating Systems Review, vol. 37, nr. 5, pag. 193-206, 2003.

[5] App Sandbox Design Guide - http://goo.gl/tjG5B

[6] B. Milewsky, Virtual Machines: Memory - http://corensic.wordpress.com/2011/11/28/virtual-machines-memory/

[7] M. Howard, Address Space Layout Randomization in Windows Vista, - http://blogs.msdn.com/b/michael_howard/archive/2006/05/26/608315.aspx

[8] File System Minifilter Drivers - http://msdn.microsoft.com/library/windows/hardware/ff540402

[9] Microsoft TechNet - Service Control Manager - http://technet.microsoft.com/en-us/library/dd349449(v=ws.10).aspx

[10] Windows Development Reference - http://msdn.microsoft.com/en-us/library/windows/desktop/hh447209(v=vs.85).aspx

[11] Windows Service Template (C++) - http://msdn.microsoft.com/en-us/library/8dy6h580(v=vs.80).aspx

[12] File System Minifilter Drivers - http://msdn.microsoft.com/library/windows/hardware/ff540402

[13] Extending Shortcut Menus - http://msdn.microsoft.com/en-us/library/windows/desktop/cc144101(v=vs.85).aspx

[14] DACLs and ACEs (Windows) - http://msdn.microsoft.com/en-us/library/windows/desktop/aa446597(v=vs.85).aspx

[15] Service Control Manager (Windows) - http://msdn.microsoft.com/en-us/library/windows/desktop/ms685150(v=vs.85).aspx

[16] Filter Manager Concepts (Windows Drivers) - http://msdn.microsoft.com/en-US/library/windows/hardware/ff541610

[17] CreateProcess function (Windows) - http://msdn.microsoft.com/en-us/library/windows/desktop/ms682425(v=vs.85).aspx

[18] How Volume Shadow Copy Service Works - http://technet.microsoft.com/en-us/library/cc785914(v=ws.10).aspx

[19] Sanboxie - Sandbox software for application isolation and secure Web browsing - http://www.sandboxie.com/

[20] George Pecherle, Cornelia Gyorödi, Robert Gyorödi, Bogdan Andronic, Iosif Ignat "New Method of Detection and Wiping of Sensitive Information", ICCP 2011, IEEE 7th International Conference on Intelligent Computer Communication and Processing, 2011, Cluj-Napoca, Romania, 25-27 August, ISBN 978-1-4577-1478-8, CFP 1109D-PRT, pages 145-148

[21] Oracle VM VirtualBox - https://www.virtualbox.org/

Redesigning Educational Systems Using IJAZA Structure

(Case study: Information Technology College)

Yasser Bahjat
Faculty of Computers and Information Technology,
King Abdulaziz University,
Jeddah, Saudi Arabia

Ibrahim Albidewi
Faculty of Computers and Information Technology,
King Abdulaziz University,
Jeddah, Saudi Arabia

***Abstract*—This paper discusses the use of IJAZA in an education system. The IJAZA system was the system used in the earliest Islamic nation for education and accreditation. The system was designed with a few major cultural concepts in mind. The basic principle of the IJAZA system is to give people complete freedom to study what they want, when they want it, from whomever they want, without any restrictions or influence from outside parties. To ensure the effectiveness of the system we think that the IJAZA system needs complete documentation of everything that happens within it, such as IJAZA documentation when it is first established, up to the performance review of the certified by the job market. Therefore, this paper provides a brief methodology, evaluation and implementation of the IJAZA system.**

Keywords—IJAZA system; Islamic educational system; Apprentice studies

I. INTRODUCTION

The IJAZA structure was the educational system used in the ancient Islamic nation for education and accreditation [1-3].The origin of IJAZA system was that Muslims keened on learning and understanding the Holy Quran and the Hadith from highly scholars. When they master what they have learned, they will obtain the IJAZA "permission" to transmit the earned knowledge in the scholars' names. With passage of days this system had been used to obtain all kind of knowledge such as math, biology, chemistry ...etc.

The IJAZA system was designed with a few major cultural concepts in mind. These concepts included: Freedom to choose what to learn, education as a desire not a mandate, knowledge should always be cited to its originator and heritage should be preserved [4,5]. Based on that, we thought that reintroducing the IJAZA system while modifying it to take into account the few centuries that have passed would be a great idea to reenergize the educational systems of today.

The description of the idea of IJAZA is as follows (Fig. 1):

a) Definition:

An IJAZA is the testimony of a scholar in a specific field that a licentiate is worthy of working in that specific field, and to give such an IJAZA to others is in many ways similar to an apprenticeship.

b) Age:

IJAZAs do not stipulate a minimum age for either the scholar or the licentiate.

c) Period:

An IJAZA is not linked to a certain period of time for completion nor is it linked to attending a minimum or maximum number of classes.

d) Worthiness:

A licentiate is given an IJAZA based on the personal evaluation by the scholar of his worthiness and is based on a specific result to a specific test.

e) Lineage:

Each IJAZA must clearly show the name of the licentiate, his scholar and the complete lineage of his IJAZA up to the founder of that IJAZA. This is the main contrast between the IJAZA system and an apprenticeship.

The wide implementation of the IJAZA system could be achieve much, such as a higher level of commitment from teachers to apprentice performance because an IJAZA is a testimony by the teacher to the licentiate, which incentivizes the teacher much more about his apprentice's performance as it would reflect directly on his personal reputation. Teachers are expected not to give an IJAZA to those who would affect his reputation negatively. Also, teacher development would improve since an IJAZA is a testimony from the scholar, and apprentices would always prefer receiving one from highly qualified scholars with strong reputations in their field. The job market would also prefer IJAZAs by scholars whose apprentices are known for their skill and professional performance (see Fig. 2), thus directly affecting the number of apprentices willing to study for his IJAZA.

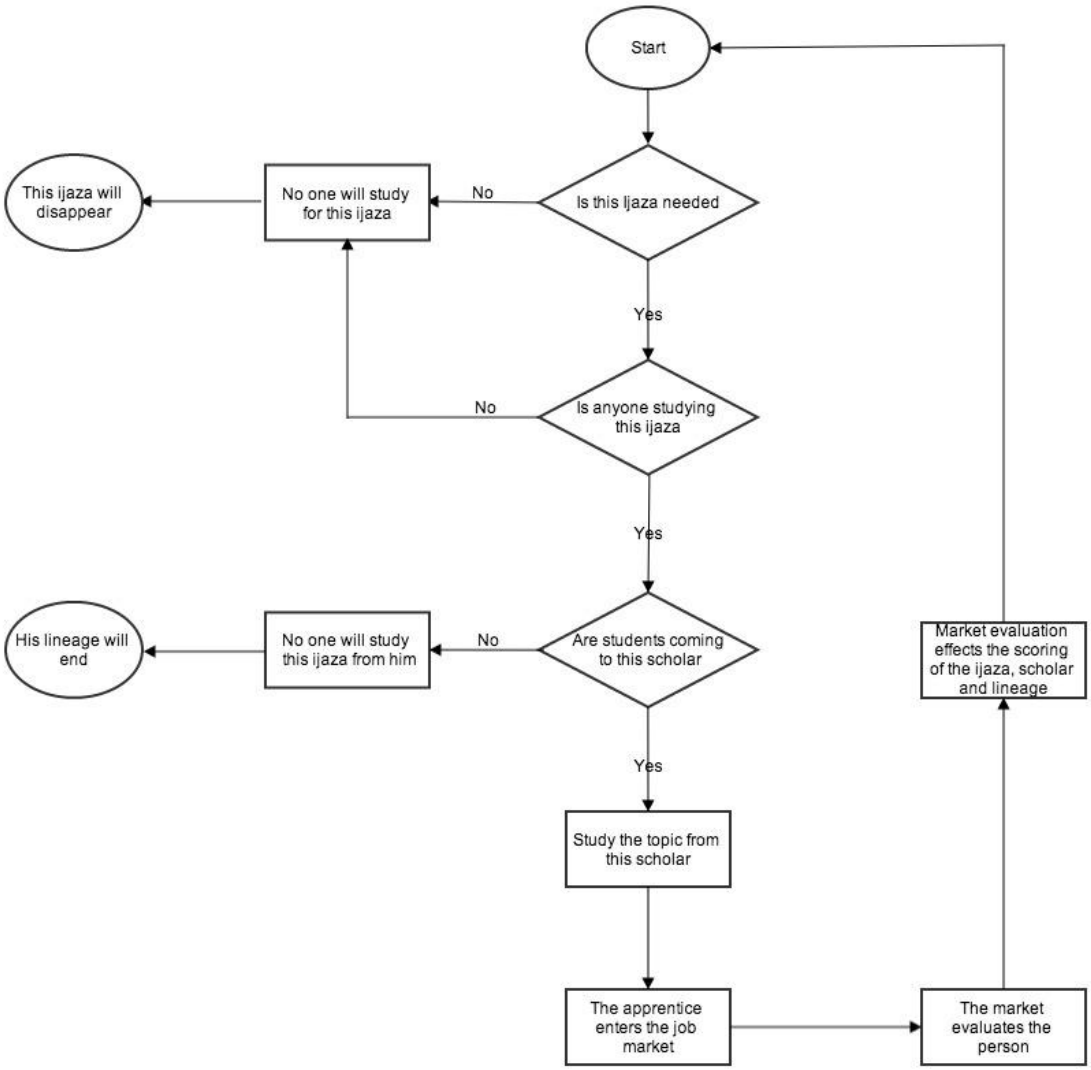


Fig. 1. IJAZA system guideline

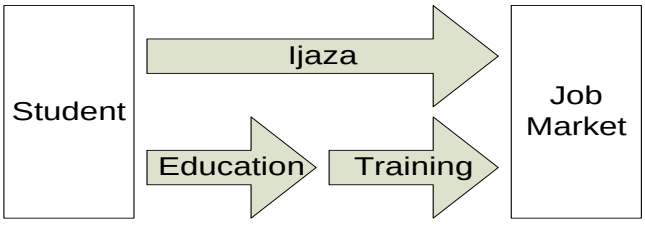


Fig. 2. Students' relation to the job market

Since an IJAZA does not depend on a specific time frame or regular attendance, the apprentices time to study and work simultaneously without either of them affecting the other, nor him needing to choose between the two since age is not a prerequisite for an IJAZA. Or an apprentice can get advanced IJAZAs at a young age allowing him to become a productive member in society instead of wasting his youth stuck a class room with zero productivity.

As the IJAZA is based solely on the discretion of the scholar, there are no restrictions on what the apprentice can or must learn from other fields. This also means that the apprentice can study from several scholars and receive the IJAZA from whomever he wants.

An IJAZA then is a true democratic system of education as the apprentice studies what he wants when he wants from whomever he wants. By allowing the apprentice to choose what he wants, the system allows each individual to be unique in a way he feels fits him, allowing him to highly specialize in any field or to be as multi-disciplined as he wants.

By giving this freedom to apprentices, they would tend to study what they are passionate about, ensuring that he would be more willing to go through hardships and have more knowledge. This would also ensure that his productivity in that field would be much higher as he is working in a field that he enjoys.

II. Literature Review

The oldest democratic education system that still exists is Summerhill, in Suffolk, England. Summerhill is a free style school. The school's philosophy is to allow freedom for the individual - each child being able totake their own path in life, and following their own interests to develop into the person that they personally feel that they are meant to be. The freedom to attend formal lessons or not at the school is a central feature of the school's philosophy [6].

Sudbury Valley School, is also a democratic education system, Sudbury Model has three basic tenets: educational freedom, democratic governance and personal responsibility. At the Sudbury Valley School, students individually decide what to do with their time, and learn as an aside to their personal efforts, interactions and ordinary experience, rather than through classes or a standard curriculum.[7]

What differs IJAZA system is that although IJAZA system is democratic system which gives people complete freedom to study what they want, when they want it, from whomever they want without any restrictions or influence, butit has a level of formality since the student must attend the classes whether on regular or irregular basis. Because IJAZA implies that the student has learned this knowledge through face-to-face interactions "at the feet" of the scholar [8]. And the only way to obtain the IJAZA is to master what you have learned and to show commitment to your studies.

III. IJAZA Methodology

The basic principle of the IJAZA system is to give people complete freedom to study what they want, when they want it, from whomever they want without any restrictions or influence from outside parties (see Fig. 3), but while providing him with access to scholars and tools as needed. It is also based on the personal responsibility of both the teacher and apprentice by linking the apprentice to his scholar and his lineage, and expecting apprentices to take responsibility of seeking out knowledge and self-improvement.

To insure the effectiveness of the system, we think that the IJAZA system needs complete documentation of everything that happens within it, such as IJAZA documentation when it is first established, up to the performance review of the certified by the job market. The following describes the main documentation of the IJAZA system:

a) Title of the IJAZA.

b) Content description of the IJAZA and its specific knowledge base.

c) Skills (topics/subjects) that the licentiate has to be aware of, and memories or skills he must master.

d) Jobs that the licentiate can work in.

e) Evaluation methods that the scholar must follow to give the IJAZA to the licentiate.

f) Requirements needed to get the IJAZA.

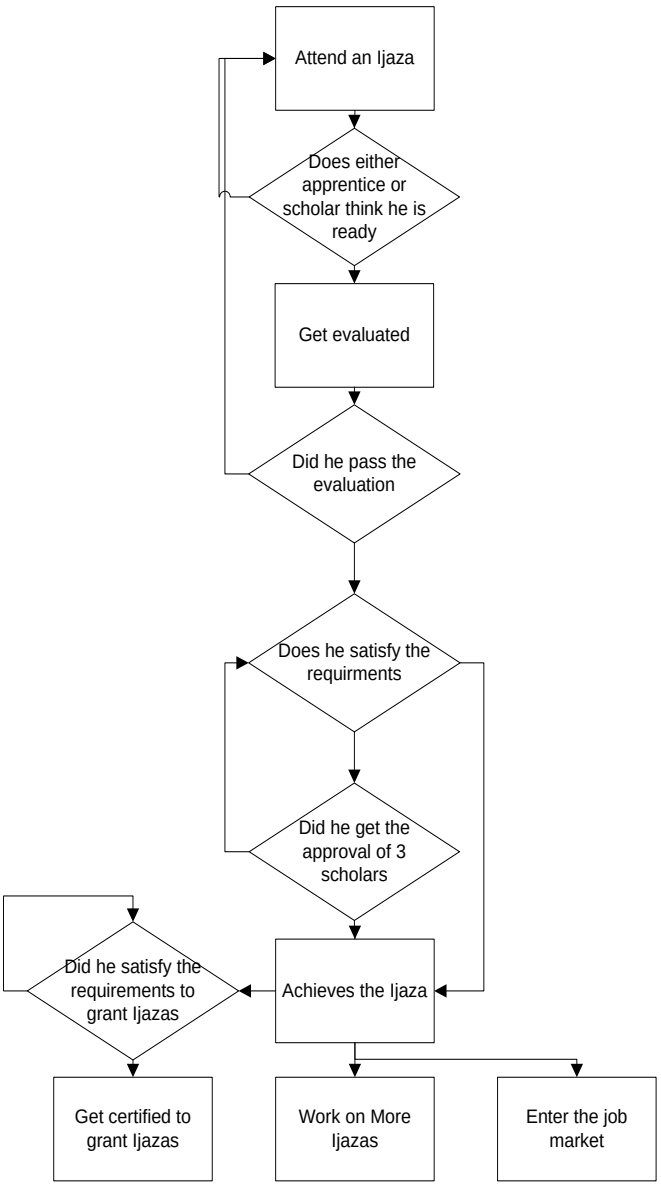


Fig. 3. IJAZA process

g) Requirements needed to be allowed to give the IJAZA to others.

h) Requirements needed to keep the IJAZA (Annual working hours, annual teaching hours, etc.).

i) Requirements and regulations for renewal of the IJAZA in the event that an IJAZA holder does not complete the requirements needed to keep his IJAZA.

The founder of any IJAZA has the right to set rules and requirements for anyone to get that IJAZA (e.g. that the apprentice has another IJAZA first, or a minimum duration of practice, a minimum duration between getting the IJAZA and him being allowed to give IJAZAs, minimum experience, etc.).

The founder also has the right to waive such requirements and so can any other subsequent scholar provided he receives the approval of three other scholars who would approve such waiving. In such case the waiver's requirements have to be listed on his IJAZA and the scholars who had approved such waiver (in such case the apprentice's lineage is only connected to his scholar if one of those who approved the waver was the founder, and in such cases he is considered connected to the founder as well as his scholar).

Any IJAZA that starts its name with the word "Basic", such as "Basic Math", such IJAZAs do not allow its holder to give an IJAZA to anyone else. The sole purpose of such an IJAZA would be to build the basic foundation for the apprentice in the field. Any IJAZA that starts its name with the word "Advanced", such as "Advanced Math", endows the licentiate to teach the "Basic" IJAZA later.

IV. IJAZA Evaluation

The evaluation philosophy in the IJAZA system is unlike other systems. As the main objective of the evaluation in the IJAZA system is to allow the scholar to understand his apprentice's capabilities and grasp on the topic, in addition to his ability to self-develop in the field. Evaluation here is based on the apprentice's request to be evaluated as he thinks he is ready to get the IJAZA, or based on the scholar's request believing that his apprentice is now ready. This means that evaluation is usually done on an individual basis for each apprentice. We see that that are four appropriate ways of evaluation in the IJAZA system [9-10], however they all end in the same way as the scholar agrees to grant the IJAZA to his apprentice based on his sole discretion:

a) Written Exams: on the surface the idea of a written exam might seem similar to the current methods of evolution in our educational and certification systems, however it is drastically different in its objectives and approach. It starts in a similar way, with the scholar writing down a few questions that the apprentice must answer in writing as well. But the similarity ends there as the way such questions are prepared and how the apprentice's answers are evaluated is different.

- Questions: since the objective is to know whether the apprentice is a master in the field on which he is being evaluated and his ability to work in the field, and since the scholar, no matter how skillful and experienced he might be, would never be able to prepare his apprentice for every possibility that he would face in his professional career. We propose that at least a quarter of the questions should be from topics not covered under the topic of the IJAZA, allowing the scholar to evaluate his apprentice's ability to adapt & innovate in addition to how his logic works when faced with things that he does not know.
- Evolution: when the scholar evaluates his apprentice's answers, he is not looking at it in the traditional manner of true and false and specific grades for each question, but as a detailed critique of those answers evaluating his methodology and style. He then discusses these comments with the apprentice and evaluates him based on his grasp of the topics covered in that exam.

b) Professional Training: as the apprentice works for a period of time in his field, his scholar (along with the person with whom the apprentice worked) evaluates the apprentice's performance.. He then discusses that evaluation with the apprentice.

c) Project: the apprentice would choose a project from an idea of his own creation, and then his scholar would publically debate his apprentice, starting with the reasons why he selected that specific project up to the final results of the project. He would also open up the floor for any of the attendants to throw in their questions on the project.

d) Verbal evaluation: it is similar to the written exam in the type of questions and how they are evaluated, however it has to be done in public. This adds new aspects for the scholar to take into consideration when making his decision.

Our recommendation is to build a database for all IJAZAs and everyone who has an IJAZA along with the complete lineage. Every time one of them takes a job his employer can evaluate his performance allowing us to build up a score for the scholar who granted him the IJAZA and giving an indication of the strength of his lineage.

The database would work as follows:

a) Registering all IJAZAs and scholars along with the complete lineage.

b) Every time someone news is granted an IJAZA, his scholar enters his information and approves his IJAZA, linking his apprentice to his lineage.

c) When anyone with an IJAZA works, his employer evaluates him on the system.

d) The system calculates the score for every scholar based on the performance of his apprentices.

e) The entire lineage chain is affected with the score change of any of its members.

f) The system would allow anyone to browse through the lineages and see how they are scored allowing the job market to predict the mastery level of new IJAZAs.

V. IJAZA Implementation

We believe that before you start teaching anything to a student/apprentice using IJAZA system, we must first teach them how to read and write and all of the skills for him to master them. Reading is not only a human's ability to know the shapes of letters and produce the proper sounds related to them, nor writing just converting sounds into a shape on a paper. Surely such skill is necessary to achieve reading and writing, but it is but one of a list of skills that a human needs to become a reader or a writer.

Reading is the method through which you receive the knowledge that the writer intended to give to you along with the feelings related to it. On the other hand writing is a way to transmit such knowledge and feelings to others by moving your own thoughts out of your head and on to the paper.

Thus we propose to establish a nightly Kuttab that does not contradict with current school timings where scholars would work with students to give them these two IJAZAs:

In this IJAZA a student would learn the shapes of letters and the proper pronunciations in addition to how to write them. A student cannot move on to the next IJAZA until he masters this one.

This IJAZA focuses on the student's ability to understand what he reads and write what he thinks and feels. The hall in which the scholar would teach his students would be a fully-fledged library with books from every field that he would use to empower his students in the different reading and writing skills that they need before he would grant them this IJAZA. Such skills include:

- Identifying the author's objectives and topic.
- Identifying the main topic and subtopics, and distinguishing between them.
- Understanding the meanings of words and structures.
- Giving a proper title to the material that they read.
- Reading properly without pronunciation errors.
- How the student interprets what he reads based on his prior knowledge and experience.
- The ability to use the dictionary to understand new words.

As previously mentioned, the paper proposes two types of IJAZA. "Basic" and "Advanced" IJAZAs. Basic IJAZA is recommended to be implemented inschools from elementary to high school and an advanced IJAZA to be implemented in higher education.

A. *Basic learning:*

You can consider this the IJAZA system's replacement for school from elementary to high school. We would recommend establishing what we call Majmaa' (Basic Science Center), where scholars would grant IJAZAs in Basic sciences in all fields such as math, biology, physics, literature, astronomy, history, etc. Students are encouraged to explore and pick whatever IJAZAs they want to achieve while providing guidance on the relationship and dependencies between them. Students are required to pass five subjects of their choice each year until they graduate.

In this way we will make sure that the students actually gained a proper education and they can contribute to science. The center must provide what the scholars might need such as books and tools to drive their teachings home and explain the practical aspects of their fields.

Initially such center would operate only in the evening not to conflict with current schooling times allowing parents to feel reassured that this new system would not negatively impact their children's futures.

B. *Advanced learning:*

You can consider this the IJAZA system's replacement for higher education and it allows students to specialize and delve deep into the knowledge areas they are passionate about. As a start, we want to establish an advanced IJAZA center for one of the fields. Students are required to have achieved one or more IJAZAs in basic sciences. This type of IJAZA will be given only to the students who master what they have learned. This type of IJAZA is equivalent to diploma and with this IJAZA students will be qualified to transmit this knowledge or to enter the job market and compete, since they have learned from professionals and those professional have certified that they are capable of working in this specific field.

This allows us to provide a higher education to those who were unable to go to college or those who achieved one of those basic IJAZAs to specialize and enter the job market, for example in the IT field as shown in Fig. 4.

VI. CONCLUSION

This paper discusses the use of IJAZA in an education system. We consider the IJAZA system as a replacement for school from elementary to high school. Therefore, it is important that this paper does not stop with the philosophical discussion of academics, to then face the harsh reality that would provide thousands of obstacles and reasons why such system cannot be implemented on a large scale. That is why we are proposing in this paper a methodology of how such a system can be implemented today, even if on a small scale where we can measure its results and provide solid data on its effectiveness.

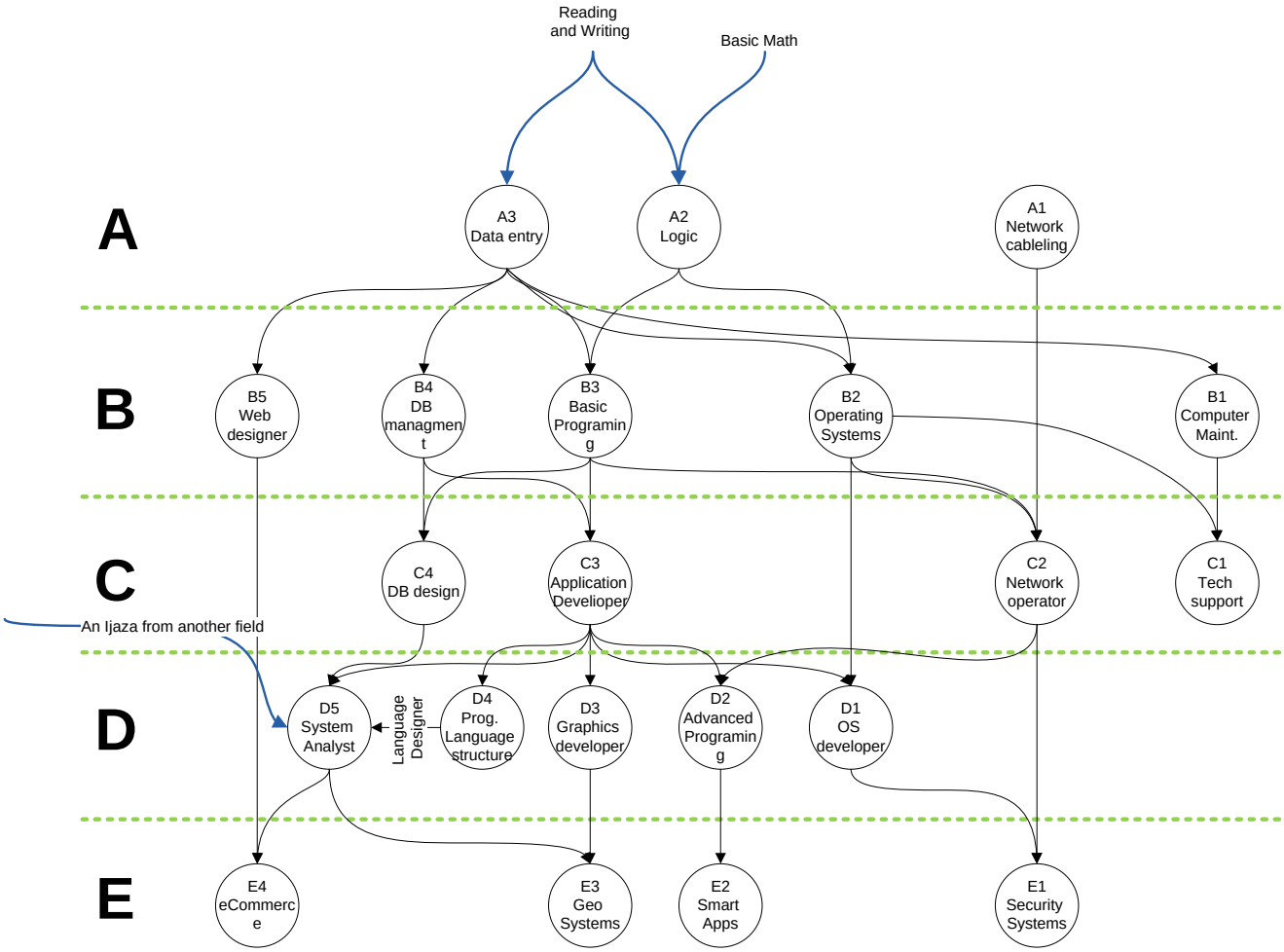


Fig. 4. Proposed Information Technology IJAZA structure

REFERENCES

[1] M. Halstead,"An Islamic concept of education," Comparative Education, vol. 40, no.4, pp.517-529, 2004.

[2] Y. Waghid, "Conceptions of Islamic education: pedagogical framings," Global Studies in Education. vol. 3, 2011.

[3] M. S. Merry, Culture, Identity, and Islamic Schooling: A Philosophical Approach. Palgrave Macmillan, 2010.

[4] A. Hassan, A. Suhid,N. Z. Abiddin, H. Ismail, &H. Hussin,"The role of Islamic philosophy of education in aspiring holistic learning," Procedia-Social and Behavioral Sciences, vol. 5, pp. 2113-2118, 2010.

[5] M. A. Lubis, M. M. Yunus, M. Diao,T. Arifin, R.M. Muhamad, N. M. Ishak,"The perception and method in teaching and learning Islamic education," International Journal of Education and Information Technologies, vol.1, no.5, pp. 69-78, 2011

[6] Summerhill. "About Summerhill." Internet: http://www.summerhillschool.co.uk/about.php, [April 14, 2014].

[7] Sudbury Valley School. "About SVS." Internet: http://www.sudval.org/01_abou_01.html, [April 14, 2014].

[8] G. William,"Tranditinalism in Islam: An Essay in Intereption," Journal of Interdisciplinary History, vol. 23, No.3, pp. 495-522, 1993.

[9] Flagg,N.Barbara. Formative evaluation for educational technologies. Routledge, 2013.

[10] Beetham, Helen, and R. Sharpe, eds.Rethinking pedagogy for a digital age: Designing for 21st century learning. routledge, 2013.

Towards a Service-Based Framework for Environmental Data Processing

Ivan Madjarov
AixMarseille Université, CNRS, ENSAM,
Université de Toulon,LSIS UMR 7296,
13397, Marseille, France

Juš Kocijan
Jozef Stefan Institute,
Department of Systems and Control, Jamova 39,
1000 Ljubljana, Slovenia University of NovaGorica,
School of Engineering and Management, Vipavska 13,5000
Nova Gorica, Slovenia

Alexandra Grancharova
BulgarianAcademy of Sciences,
Institute of System Engineering and Robotics,
P.O.Box 79, Sofia 1113,
Bulgaria

Bogdan Shishedjiev
TechnicalUniversity of Sofia,
Department Programming and Computer Technologies,
1000 Sofia Bulgaria

***Abstract*—Scientists are confronted with significant data management problems due to the huge volume and high complexity of environmental data. An important aspect of environmental data management is that data, needed for a process, are not always in the adequate format. In this contribution, we analyze environmental data structure, and model this data using a semantic-based method. Using this model, we design and implement a framework based on Web services for transformation between massive environmental text-based data and relational databases. We present a mapping model for environmental data transformation to be used in the scenario devoted to the methodology for development of stochastic models for prediction of environmental parameters by application of Gaussian processes.**

Keywords*— *Scientific data; Environmental data; Web services; Data integration; Stochastic model; Gaussian process; Metadata

I. INTRODUCTION

Nowadays, a significant part of a scientist's work is dedicated to accessing, visualizing, integrating and analyzing data from a wide range of heterogeneous sources because science is more and more data-driven. On the other hand, scientist's activities, scientific instruments and computer simulations produce more and more data from different domain, e.g. physics, astronomy, meteorology, air pollution and so on. Scientists process these data and generate new data based on the results of the processes. Editing and updating of data also generates data. Produced data are schema-less, semi or fully structured persisting in different repositories [5]. According to some sources [2], the data volumes are approximately doubling each year. Furthermore, scientists need to know based on which collection of data they have produced a specific result. An important problem that arises here is the data provenance and the data versioning that can be expressed by the question: What data in which version a specific result was obtained. So, data require new methods of organization for scientific analysis. It is obvious that scientists need a data structuring and a storing organization for data management and processing.

The existing scientific tools are mostly focused on data processing and visualization, and data management is largely left to the user [3].

Many of scientific data are traditionally stored in ASCII format, i.e. text file. The ASCII text is a recognized standard for data exchange (e.g. input/output) supported by scientific instruments and simulation devices. It is recognized that ASCII-based data are platform independent, so they can be analyzed in different operating systems and they can be imported to whatever information system or scientific workflow. However, this form presents some drawbacks:

- *Low readability*: data can be presented in different units without any context-based explanation and they become somewhere ambiguous.
- *Hard to integrate*: scientific data are natively heterogeneous, unstructured and they are usually stored in different files and/or in different locations. This makes it difficult to integrate all the data into one place without a common sematic schema.
- *Data searching*: content discovery is a difficult task in a large datasets or in thousands of distributed files.

An important aspect of environmental data management is that data, needed for a process, are not always in the adequate format. Scientists use different tools in different stages of their research; they develop some tools for their work by themselves and spend time to retrofitting data into acceptable formats for these tools [4].

So, the main problem to address here is how to provide an efficient way to implement massive data transformation between texts and databases. This is a common problem for both computer science researchers and environmental science researchers, aswe consider environmental data as a subset of scientific data.

In semi-structured data, the information that is normally associated with a schema is contained within the data [3]. The

meaning and logic structure of semi-structured data can be expressed and identified by semantic tags. For instance, XML is a standardized extended markup semi-structured data.

In this paper, we present our work in progress. We analyze environmental data structure, and model this data using a semantic-based method. Using this method, we design and implement a Web service-based framework for transformation between massive environmental text-based data and relational databases. As main contribution, we present a mapping model for environmental data transformation. We apply this model in a scenario devoted to the methodology for development of stochastic models for prediction of environmental parameters.

We envision a schema for prediction of environmental parameters by application of Gaussian processes, e.g. the ozone concentration in the air based on data collected on-line by automatic measurement stations. As well, we can easy apply the developed methodology to predict the concentrations of other air pollutants e.g. sulfur dioxide and nitrogen dioxide.

The paper is organized as follows: first in section 2 we present the background with some related work. In section 3, we present our motivation and concept for an environmental Web services-based workflow. In sections 4 and 5 we present our scenario for environmental data processing based on Gaussian processes. Finally, in section 6 we conclude and discuss some future work.

II. Background And Related Work

As presented in [3] scientific utilities can fall into three categories: (1) scientific software; (2) scientific languages and (3) scientific workflows. In this study we present a non-exhaustive list of mature scientific utilities i.e. scientific software, scientific languages, workflows software and systems to justify the choice that we will do in our research project.

A. Scientific Software

Scientific software tools in general, load data in memory. Usually scientists need to perform some extra steps in order to prepare data for processes. To use different tools, scientists must learn different sets of commands, scripting or programming languages for different framework and operating systems.

B. Scientific Languages

The *Apache Hadoop* [9] is an open-source software library for storage and large-scale processing of data-sets on clusters. It is a framework that allows distributed processing of large data sets across single servers or thousands of machines by using simple programming models. As presented in [8], the Java open source library is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster.

Google Open Refine (GOR) [6] is a standalone open source desktop application for data cleanup and transformation to other formats. It is similar to spreadsheet applications; however, it behaves more like a database. GOR opens a Web interface powered by aWeb server. It operates on rows of data which have cells under columns, which is similar to relational database tables. Transformation expressions are written in the GOR Expression Language. It is able to work with CSV, TSV, XML, JSON, Excel and RDF formats.

Matlab [7] is a numerical computing environment and allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces. It imports data from CSV files, excel spreadsheets and databases. Import functions read the data in memory and reorganize them in vectors or matrices, then all functions work on these data structures and possible interfacing with programs written in other languages, including C, C++, Java, and FORTRAN.

C. Scientific Workflows

Workflow composition represents the conceptual model of a scientific analysis which implies the flow of data within a system. Every step of workflows acts on the data. The required data are obtained from previous steps, from local files, from relational databases, from remote services or another source.

Kepler [10] is a free scientific workflow management system. It is able to acquire data from different sources, process them by prepared or user defined components. Optionally, an external data processing facility can be applied. This software provides process and data monitoring, provenance information, and data movement solutions. Its architecture is directed graphs where the nodes represent discrete computational components and the edges represent paths along which data and results can flow between components. In Kepler obtaining data from external sources like CSV files, spreadsheets, relational DBMSs and remote data sources are done by specific actors as metaphors to model the steps of workflows. The system includes a graphical user interface for composing workflows.

VisTrails [11] is an open-source scientific workflow and provenance management system that provides support for simulations, data exploration and visualization [3]. The provenance information is presented as XML files or as tables in a relational database. It allows users to navigate workflow versions, to undo changes, to visually compare workflows and their results, and to examine the actions that led to a result. It allows the combination of loosely-coupled resources, specialized libraries, grid and Web services.

Taverna [12] is an open source scientific Workflow management tool suite to design and execute workflows. It is able to fetch data from CSV and spreadsheet files, local and remote resources through provided or custom services. It provides provenance functionalities and a common model for workflows and means for sharing and reusing them across the borders of individual working groups. To leverage the existing infrastructure, the computational model strongly focuses on Web-services. It provides an API and a Web interface to access data about various Web services.

III. Motivation and Concept for an Environmental Workflow

Scientific Workflows present a managed combination of activities and computations in order to resolve scientific problems. In contrast to business Workflows that implement business processes involving different actors and systems, scientific workflows are used to realize computational experiments, possibly confirming or invalidating scientific

hypotheses. Scientific Workflow systems maintain the execution of repetitive tasks such as data access, transformation and analysis [1, 24] data from heterogeneous sources, e.g. sensor systems, measuring instruments, text files, spreadsheets, databases, simulation devices, etc.

The creation and exchange of scientific and environmental information increase the amount of data that should be processed, from one hand, as well as the possibilities for their interpretation, on the other hand. This motivates many researchers and specialists to reconsider the existing engineering and network architectures, the database schemas, the algorithms and rules for data interpretation. Beside the hug size, the data are represented in a way, which does not allow processing by the traditional DBMS, because of their heterogeneity and specific characteristics.

Sensor systems are usually used to monitor the state of the environment in the urban areas. The obtained measurements need to be stored in a database, which is very important for the development of schema-based data models. So, the data collected by the sensors are used in real time by different applications through procedures for control of large amount of data in spatiotemporal databases. The problem which arises is related to the information control, because of the specific characteristics of the collected data. The space-time character of data requires the development of new approaches for structuring, exploitation and visualization of these data. Sensor networks and associated databases are used for monitoring and registration of various environmental phenomena, e.g. for the accurate prediction of the future values of these phenomena and for all stochastic-based data processing for environmental norm evaluation.

Specific languages for scientific data description already exist. CDF and HDF are languages, which are used in the physics of thermonuclear synthesis, the geology and the astronomy. They represent data models, API, and file formats for storage and control of scientific data. These formats allow storing data as a simple table that is difficult to apply with a large amount of data that have a complex structure. In our work we process environmental data as a subset of scientific data. However, a specific language for description of environmental data doesn't really exist. Moreover, there is a large diversity of characteristics proprietary of environmental data, i.e. different scales of measurement expressed in different units. We suggest the use or the extension of a scientific Workflow with adapted semantics for presentation and storage of large amount of data, related to the monitoring system that analyzes environmental parameters. We argue for a semantic and Web service-based approach for processing environmental data from multiple and heterogeneous sources.

The study of environmental data requires the use of protocols, mathematical models and procedures, which need to be validated. In order to accomplish this, we rely on a Workflow scientific process through integration and control of the components, defining the air quality in the environment.

The scientific goal of our research work in progress is to study the complexity of the systems for environmental monitoring, which use large amount of data.

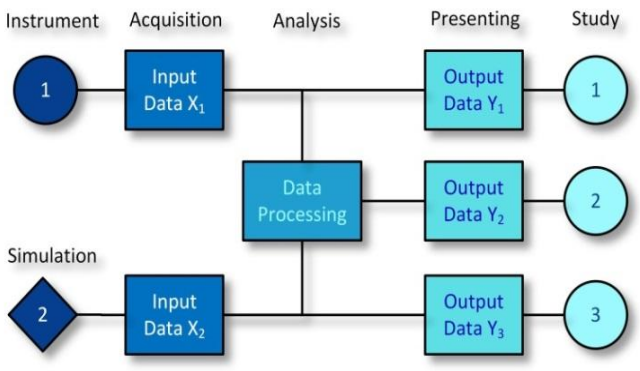


Fig. 1. General structure diagram of a scientific workflow for environmental data processing.

We develop solutions in terms of semantic languages, models and methods for access, storage and use of scientific and environmental data, implemented mainly as Web interfaces and services. Our focus in this area is geared towards the design and implementation of service oriented systems that allow a pay-as-to-go generation of composite cloud-based services according to the users' requirements.

In this paper, we aim at the development and integration of technologies and expertise, necessary to resolve the problem with the huge amount of environmental data applied to stochastic models for environmental parameter prediction by application of Gaussian processes. In order to achieve this goal, we rely on a Workflow-based scientific process, directed towards the control of data flow (Fig. 1).

The main goal includes the following sub-goals:

1) Development of a data control strategy. We study the algorithms and the efficiency of the Cloud-XaaS platform (Anything-as-a-Service) with an emphasis on the semantic structuring of acquired data from the instruments in order to facilitate the data integration when heterogeneous sources are used.We develop services for remote data control, associated with the data processing,i.e.acquisition, analysis, requests, actualization,computations and visualizationas shown inFig. 1.

2) Data storage. We develop a multi-layer model with an automatic indexing of data by using the existing services within the Cloud-based platforms. We propose a native data storage architecture (NXD), which is adapted to various functions allowing the connection with other platforms.

3) Distribution of the environmental data. We develop a model for digital visualization of environmental data through a transformation process for Web-based presentation in terms of tables and/or vector graphics. The environmental data are transformed into SVG, as an XML document, which allows building applications for immediate graphical representation of the prognosis on the user side. The digital visualization is associated with the latest advances in responsive design that takes into account all particularities of desktop and mobile devices based on media queries.

4) Development of mathematical models for prediction of environmental parameters. This includes the system

integration via Web services of the modeling approach based on Gaussian processes with data about the concentrations of ozone, sulfur dioxide and nitrogen dioxide in the air, collected at the automatic measurement stations.

5) Metadata descriptions. Scientific Workflow systems typically describe data processing via a Workflow definition language. However, current specific Workflow definition languages, even adopted by current mature scientific Workflow systems, are too complex and excessive for non-professionals. We design an XML-based environmental data definition language using schema descriptions to suit a lightweight workflow system in a specific domain such as air quality.

6) Data integration. Notable characteristics of scientific computing are data integration, data manipulations during calculation, scientific analysis, data migration and the data store on distributed machines according to guidelines and logical relations [8]. We assert that Web Services can be used to unlock heterogeneous scientific systems to extract and integrate environmental data.

There are two issues in using a scientific Workflow approach to prediction modeling: The first one is the choice of a Workflow composition and execution environment. The second issue adapts the process steps in an environmental data management suite. We recommend the second issue because it can be associated to Web Service technology. It is necessary to recall that the Web Service paradigm enables the aggregation of multiple data sources. In this approach, each process step is implemented as a Web Service and Web Services are chained together to form a modeling task as shown in Fig. 2. In the core of Web Service technology is the Web Services Description Language (WSDL) [13]. WSDL provides a XML-based framework and language for defining interfaces e.g. input and output, SOAP access specification (Simple Object Access Protocol) [14] and the location of the service. This approach can achieve greater system interoperability with existing scientific Workflows.

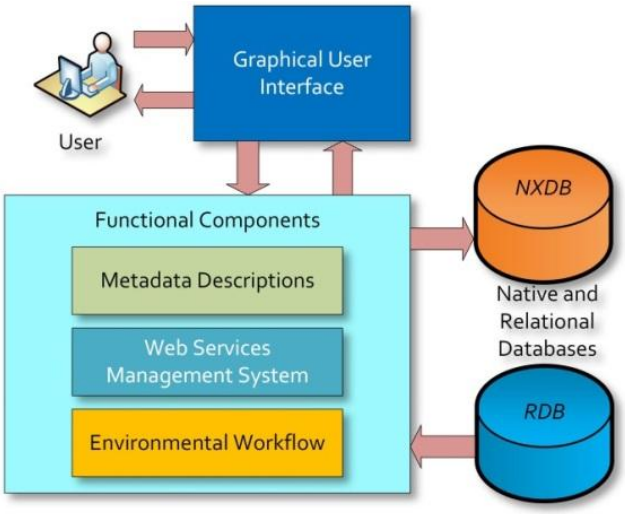


Fig. 2. Integrated platform for environmental data processing with environmental metadata description, Web Services Management for data and application integration with an environmental Workflow.

IV. Environmental Data Metadata Description and Processing

A. Environmental Data Need a MetadataDescription

In general, to be able to process scientific and environmental data it is important to know their meaning, e.g. what it is about, how they was obtained, how they are formatted and so on. This information is coded and stored as data about the real data, i.e. an underlying definition or description. The formal descriptions are useful to record meaningful information about the data, their provenance and their coding in order to be understood by other users. So, we generate metadata as data that describe other data with some common characteristics:

- The metadata summarize basic information about data, which can make finding and working with particular instances of data easier, or to locate a specific set of data by filtering through metadata.
- Metadata for scientific and environmental data contain descriptions of the content, as well as keywords linked to the content. These are usually expressed in the form of meta-tags.
- The meta-tags are the vocabulary of metadata and they are often evaluated by search engines to help decide of data relevance.
- The metadata information is to be used in automated data processing by standard procedures, i.e. the procedures have to understand metadata and to process data according to metadata description.
- Metadata can be created manually, or by automated information processing.

There are a lot of research works in the metadata domain as described in [24].Some of them try to define a formal language able to describe a widest set of data. Organization such as OMG[23] developed standard models and languages such as CWM and UML. On the basis of CWM several metadata models for business application were developed in [22]. The main difficulty to address here is the data heterogeneity, the variety of their applications and the wide range of specialized languages used for their description. The native heterogeneity specific to environmental data requires a meta-description that takes into account the difference in size, the difference in measurement scale, the difference in context or provenance. In this study, we find that languages mentioned so far do not appear to be entirely satisfactory. Therefore we recommend more appropriate environmental data semantics to be defined.

B. Metadata Types and Models

In our research work for the description of environmental data we define different types (levels) of metadata:

*1) Origin:*this data describe the ownership of each piece of data, the place where it is stored, the organization and/or the person responsible for its maintenance.

*2) Access right:*this data describe rights to read, write or process data by someone.

*3) Processing:*the data about special routines or/and algorithms for processing a piece of data.

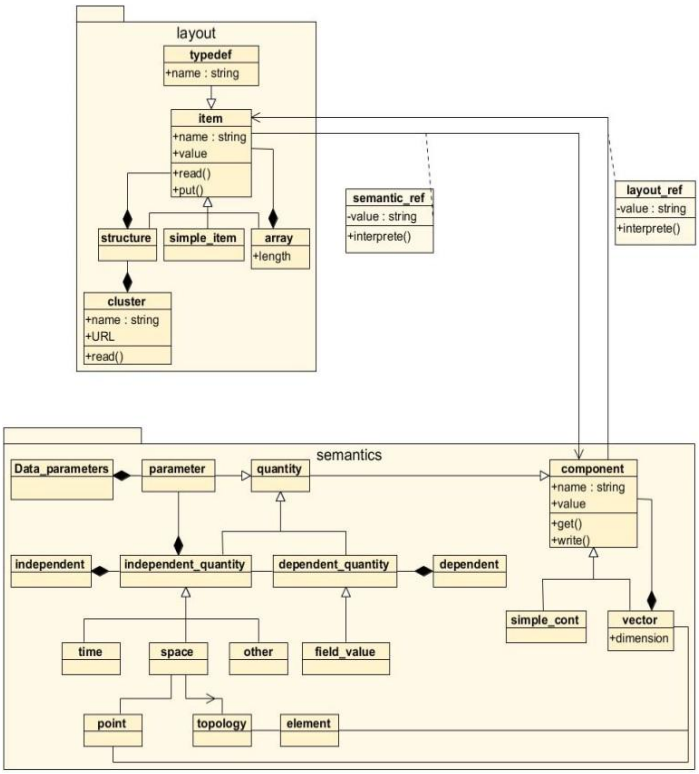


Fig. 3. Environmental data model presented in three sections with two reference classes that tell the mapping between both models presented as XML schema, i.e. (1) General; (2) Semantics; (3) Layout.

*4) Formatting:*this data describe how data are recorded and stored; are they numerical and in what unit of measurement are they written.

5) Naming and Meaning: this data describe data about the namespace of every piece of data, their meaning described by the language of the knowledge domain and the data provenance.

Fig. 3 shows our concept of XML schema for metadata description. This choice is argued by the differences between business data and scientific data as described below:

- Most scientific data is numerical and float especially in domains with strong mathematical background as physics, chemistry and engineering.
- The datasets concerning one source are huge.
- The origin and access metadata values are identical for whole datasets. They do not differ from value to value as in the case of business data.
- Most of scientific data are multidimensional tables.

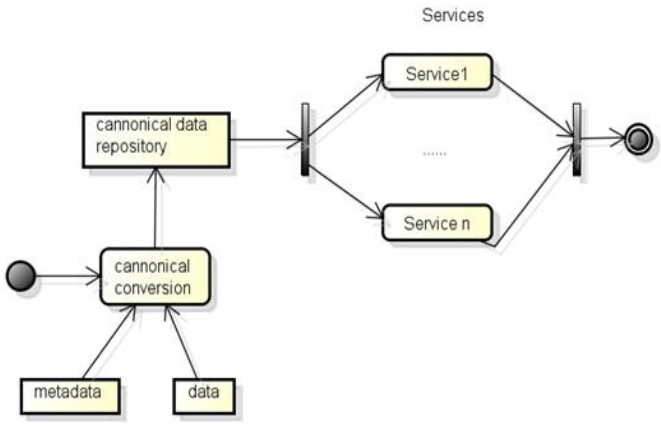


Fig. 4. Canonical form of our environmental data processing model

C. Relational Model of Scientific Data

In our work we propose a scientific-environmental dataset as a simple relational database. So, the metadata was devised in three parts (sections).

1) General: metadata about the origin and access rights. This part contains also a general description of the data and references to specific procedures used to process them.

2) Semantics: contains elements describing the meaning of the data file. The main hypothesis was that most environmental data can be presented as one or as few tables containing two types of quantities: (1) independent and (2) dependent quantities. This way they can be examined as a relational table with a primary key consisting of independent quantities and the dependent quantities as non-key attributes. There are other data named parameters that are common for the whole dataset and characterize the environment of the experiment or the assumptions of the simulation as shown in Fig. 3.

*3) Layout:*describes the formatting and the structure of the raw data.

In Fig. 4 our environmental data processing model is presented. The idea behind is to convert environmental data to the structure of the developed semantics model named canonical or standard form. By this approach it becomes easier to develop associated Web services for environmental data processing. Instead developing $M \times N$ different Web services processing M different data structures to N results we can produce M transformations (automatic) to standard form and N Web services. The conversion is done according the meta-description of data and Web services defined in the canonical description shown in Fig. 3.

The proposed solution serves as a modeling language for experimental and measured data from different environmental sources and captures, especially applied to predict the concentrations of air pollutants in an inspected region.

V. Environmental Data Processing Based On Gaussian Processes

This section is devoted to the methodology for development of stochastic models for prediction of environmental parameters by application of Gaussian processes. It represents the core of the data processing block in the structural diagram, shown in Fig. 1. The Gaussian process (GP) model is a probabilistic, non-parametric black-box model based on the principles of Bayesian probability. The output of the GP model is a random variable withnormal distribution, expressed in terms of the mean and the variance. The mean value represents the most likely output and the variance can be interpreted as a measure of its confidence. The obtained variance, which depends on the amount and the quality of the available identification data, is important information when it comes to distinguishing the GP models from other computational intelligence methods. Because of their properties GP models are especially suitable for uncertain processes modelling or when modelling data are unreliable, noisy or missing. In this respect, GP models fit well for environmental systemmodelling. Its use and properties for modelling are reviewed in [15]. The use of Gaussian processes for dynamic system modelling is a relatively recent development [16, 17, 18]. A retrospective review of dynamic system modeling with Gaussian process models can be found in [19].

A Gaussian process is a collection of random variables which have a joint multivariate Gaussian distribution (Fig. 5). Assuming a relationship of the form $y = f(\mathbf{x})$ between an input $X \in R^D$ and output $Y \in R$, we have $y(1), y(2), \ldots, y(M) \sim \mathrm{N}(0, \boldsymbol{\Sigma})$, where $\Sigma_{pq} = \mathrm{Cov}(y(p), y(q)) = C(\mathbf{x}(p), \mathbf{x}(q))$ gives the covariance between the output points $y(p)$ and $y(q)$ corresponding to the input points $\mathbf{x}(p)$ and $\mathbf{x}(q)$. Thus, the mean $\mu(\mathbf{x})$ (usually assumed to be zero) and the covariance function $C(\mathbf{x}(p), \mathbf{x}(q))$ fully specify the Gaussian process. Note that the covariance function $C(\mathbf{x}(p), \mathbf{x}(q))$ can be any function with the property that it generates a positive definite covariance matrix. A common choice is:

$$C(\mathbf{x}(p), \mathbf{x}(q)) = v_1 \exp\left[-\frac{1}{2}\sum_{i=1}^{D} w_i (x_i(p) - x_i(q))^2\right] + v_0 \alpha_{pq} \qquad (1)$$

where $\boldsymbol{\Theta} = [w_1, \ldots, w_D, v_0, v_1]$ are the "hyper-parameters" of the covariance function, x_i denotes the *i*-th component ofthe *D*-dimensional input vector$\mathbf{X}$, and α_{pq} is the Kronecker operator. The covariance function (1) is composed of two parts: the Gaussian covariance function for the modeling of system function and the covariance function for the modelling of noise. The noise is usually presumed to be white. Other forms of covariance functions suitable for different applications can be found in [15]. For a given problem, the hyper-parameters are learned (identified) using the data at hand. After the learning, one can use the w parameters as indicators of 'how important' the corresponding input components (dimensions) are: if w_i is zero or near zero it means that the inputs in dimension *i* contain little information and could possibly be removed.

Consider a set of *MD*-dimensional input vectors $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M]^T$ and a vector of output data $\mathbf{y} = [y_1, y_2, \ldots, y_M]^T$. Based on the data $(\mathbf{X}, \mathbf{y})$, and given a new input vector $\mathbf{x}^*$, we wish to estimate the probability distribution of the corresponding output y^*. Unlike other models, there is no model parameter determination as such, within a fixed model structure. With this model, most of the effort consists in *tuning* the parameters of the covariance function. This is done by maximizing the log-likelihood of the parameters, which is computationally relatively demanding since the inverse of the data covariance matrix (*M*×*M*) has to be calculated at every iteration.

The described approach can be easily utilized for regression calculation. Based on a training set $\mathbf{X}$, a covariance matrix $\mathbf{K}$ of size *M*×*M* is determined. As already mentioned before, the aim is to estimate the probability distribution of the corresponding output y^* at some new input vector $\mathbf{x}^*$. For a new test input $\mathbf{x}^*$, the predictive distribution of the corresponding output is $y^* \mid \mathbf{x}^*, (\mathbf{X}, \mathbf{y})$ and is Gaussian, with mean and variance:

$$\begin{aligned} \mu(\mathbf{x}^*) &= \mathbf{k}(\mathbf{x}^*)^T \mathbf{K}^{-1} \mathbf{y} \\ \sigma^2(\mathbf{x}^*) &= k_0(\mathbf{x}^*) - \mathbf{k}(\mathbf{x}^*)^T \mathbf{K}^{-1} \mathbf{k}(\mathbf{x}^*) \end{aligned} \qquad (2)$$

where $\mathbf{k}(\mathbf{x}^*) = [C(\mathbf{x}_1, \mathbf{x}^*), \ldots, C(\mathbf{x}_M, \mathbf{x}^*)]^T$ is the *M*×*1* vector of covariance between the test and training cases and $k_0(\mathbf{x}^*) = C(\mathbf{x}^*, \mathbf{x}^*)$ is the covariance between the test input and itself.

The identified model, in addition to mean value, also provides information about the confidence in prediction by the variance. Usually, the prediction confidence is depicted with 2σ interval which is about 95% confidence interval.

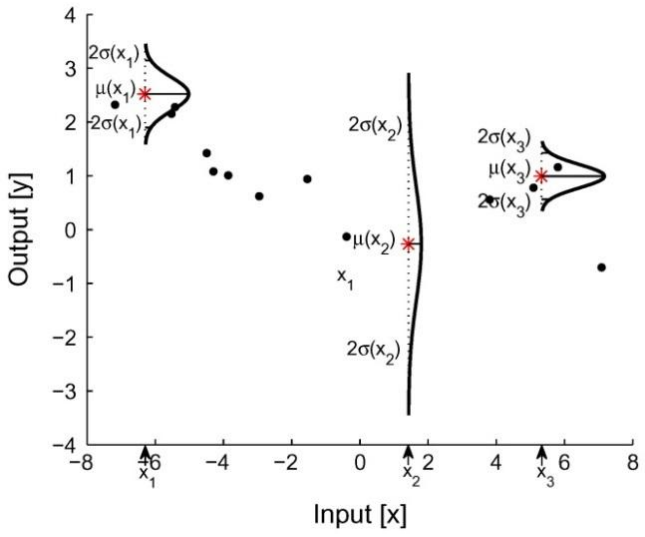


Fig. 5. Modelling with GP - Gaussian distribution of predictions at new points x_1, x_2 and x_3, conditioned on the training points (.).

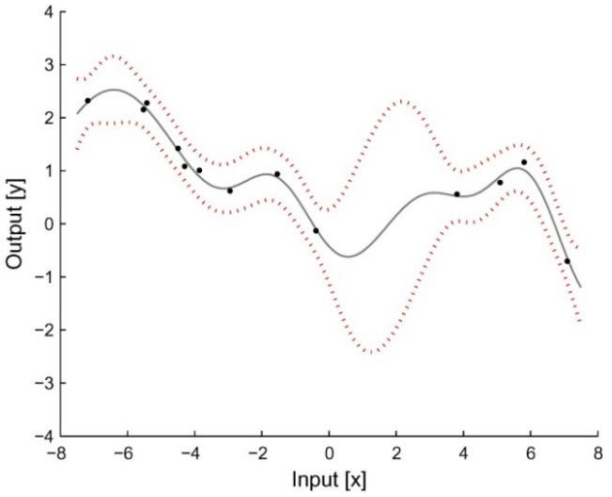


Fig. 6. Using GP models - in addition to the prediction mean value (full line), we obtain a 95% confidence region (dotted lines) for the underlying function f .

This confidence region can be seen in the example in Fig. 6as band bounded by dotted lines. It highlights areas of the input space where the prediction quality is poor, due to the lack of data or noisy data, by indicating a wider confidence band around the predicted mean.

Gaussian processes can, like neural networks, be used to model static non-linearities and can therefore be used for modeling dynamic systems [16, 18] as well as time series if lagged samples of output signals are fed back and used as regressors. For the environmental parameter dynamics modelling we consider representation where the output at the step *k* depends on the delayed outputs *y*:

$$y(k) = f(y(k-1), y(k-2), \ldots, y(k-n)) + \varepsilon(k) \quad (3)$$

where $\varepsilon(k)$ is white noise and the output $y(k)$ depends on the vector $[y(k-1), y(k-2), \ldots, y(k-n)]^T$. Assuming the signal is known up to the step *k*, we wish to predict the system output *h* steps ahead, i.e., we need to find the predictive distribution of $y(k+h)$ corresponding to $\mathbf{x}(k+h)$. Multiple-step-ahead predictions of a system modeled by eq. (3) can be achieved by iteratively making repeated one-step-ahead predictions, up to the desired horizon [16, 18].

The quality of the mean values predicted by a Gaussian process model can be assessed by computing the average squared error (ASE) [15]:

$$ASE = \frac{1}{M}\sum_{i=1}^{M}[\hat{y}_i - y_i]^2 \quad (4)$$

where $\hat{y}_i$ and y_i are the output prediction and the output measurement at the *i*-th step. Additionally, the quality of the prediction variance can be assessed with the logarithm of the predictive density error (LD) [15]:

$$LD = \frac{1}{2}\log(2\pi) + \frac{1}{2M}\sum_{i=1}^{M}\left(\log(\sigma_i^2) + \frac{[\hat{y}_i - y_i]^2}{\sigma_i^2}\right) \quad (5)$$

where σ_i^2 are the prediction at the *i*-th step.

The described methodology for development of GP models for environmental parameter prediction has been already applied to predict the ozone concentration in the air of Bourgas city, based on data collected on-line by the automatic measurement stations [20, 21]. This methodology can be easily applied to predict the concentrations of other air pollutants like sulfur dioxide and nitrogen dioxide in some of the most air polluted industrial cities in Bulgaria (Plovdiv, Stara Zagora, Varna, Bourgas).

VI. Conclusion

In this paper we proposed a concept of the framework for environmental data processing and stochastic models for prediction of environmental parameters. We analyzed environmental data structure, and modeled this data using a semantic-based method. Using this model, we designed and implemented a framework based on Web services for transformation between massive environmental text-based data and relational databases. We presented a mapping model for environmental data transformation to be used by application of Gaussian processes.

For future work we emphasize for environmental risk management and data provenance linked to gas emissions and pollution of air in industrialized cities.

References

[1] Joost N. Kok, Anna-Lena Lamprecht, and Mark D. Wilkinson, Tools in Scientific Workflow Composition, T. Margaria and B. Steffen (Eds.): ISoLA 2010, Part I, LNCS 6415, pp. 258–260, Springer, 2010.

[2] Shi Feng, Jie Song, Xuhui Bai, Daling Wang, and Ge Yu, A Web-Based Transformation System for Massive Scientific Data, L. Feng et al. (Eds.): WISE Workshops, LNCS 4256, pp. 104 – 114, Springer, 2006.

[3] Javad Chamanara, Birgitta König-Ries, SciQL: A Query Language for Unified Scientific Data Processing and Management, In: PIKM'12, Maui, Hawaii, USA, 2012.

[4] P. Prabhu, T. B. Jablin, A. Raman, Y. Zhang, J. Huang, H. Kim, N. P. Johnson, F. Liu, S. Ghosh, S. Beard, T. Oh, M. Zoufaly, D. Walker, D. I. August. A Survey of the Practice of Computational Science. In ACM, editor, State of the Practice Reports, pp. 19:1–19:12, ACM Press, 2011.

[5] A. Ailamaki, V. Kantere, and D. Dash. Managing scientific data. Commun. ACM, 53(6):68–78, 2010.

[6] Google Refine. http://code.google.com/p/google-refine/.

[7] Matlab: The Language of Technical Computing. http://www.mathworks.com/products/ matlab/.

[8] Gaozhao Chen, Shaochun Wu, Rongrong Gu, Yongquan Xu, Lingyu Xu, Yunwen Ge, Cuicui Song, Data Prefetching for Scientific Workflow Based on Hadoop, Computer and Information Science 2012, Studies in Computational Intelligence Volume 429, pp 81-92, Springer, 2012.

[9] Apache Hadoop , http://hadoop.apache.org/

[10] The Kepler Project. https://kepler-project.org.

[11] VisTrails. http://www.vistrails.org/index.php/Main_Page.

[12] Taverna Workflow Management System. http://www.taverna.org.uk.

[13] W3C, WSDL, http://www.w3.org/TR/wsdl.

[14] W3C, SOAP, http://www.w3.org/TR/soap/.

[15] C. E. Rasmussen and C. K. I. Williams. Gaussian processes for machine learning, MIT Press, Cambridge, MA, London, 2006.

[16] K. Ažman, J. Kocijan. Application of Gaussian processes for black-box modelling of biosystems. ISA Transactions, Vol. 46, No 4, pp. 443-457, 2007.

[17] A. Grancharova, J. Kocijan, and T. A. Johansen. Explicit stochastic predictive control of combustion plants based on Gaussian process models. Automatica, vol. 44, No. 6, pp. 1621-1631, 2008.

[18] J. Kocijan, A. Girard, B. Banko, R. Murray-Smith. Dynamic systems identification with Gaussian processes. Mathematical and Computer Modelling of Dynamic Systems, vol. 11, No. 4, pp. 411-424, 2005.

[19] J. Kocijan, Dynamic GP models: an overview and recent developments, ASM'12 Proceedings of the 6th international conference on Applied Mathematics, Simulation, Modelling, Pages 38-43, 2012.

[20] D. Petelin, J. Kocijan, and A. Grancharova. On-line Gaussian process model for the prediction of the ozone concentration in the air. Proceedings of BAS, vol.64, No.1, pp.117-124, 2011.

[21] D. Petelin, A. Grancharova, and J. Kocijan. Evolving Gaussian process models for prediction of ozone concentration in the air. Simulation Modelling Practice and Theory, vol.33, pp.68-80, 2013.

[22] David Marco, Michael Jennings, Universal MetaData Models, ISBN 0-471-08177-9, Wiley 2004.

[23] OMG Specifications, http://www.omg.org/spec/.

[24] A. Ailamaki, V. Kantere, D. Dash, Managing Scientific Data, Communications of the ACM, vol. 53, 6, pp.68-78, 2010.

Improving the Prediction Accuracy of Multicriteria Collaborative Filtering by Combination Algorithms

Wiranto
Informatics Department
Sebelas Maret University
Surakarta, Indonesia

Edi Winarko
Department of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia

Sri Hartati
Department of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia

Retantyo Wardoyo
Department of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia

***Abstract*—This study focuses on developing the multicriteria collaborative filtering algorithm for improving the prediction accuracy. The approaches applied were user-item multirating matrix decomposition, the measurement of user similarity using cosine formula and multidimensional distance, individual criteria weight calculation, and rating prediction for the overall criteria by a combination approach. Results of the study show variation in multicriteria collaborative filtering algorithm, which was used for improving the document recommender system, with the two following characteristics- first, the rating prediction for four individual criteria using collaborative filtering algorithm by a cosine-based user similarity and a multidimensional distance-based user similarity; second, the rating prediction for the overall criteria using combination algorithms. Based on the results of testing, it can be concluded that a variety of models developed for the multicriteria collaborative filtering systems had much better prediction accuracy than for the classic collaborative filtering, which was characterized by the increasingly smaller values of Mean Absolute Error. The best accuracy was achieved by the multicriteria collaborative filtering system with multidimensional distance-based similarity.**

Keywords*—Algorithm; multicriteria collaborative filtering; document; recommendation; system; similarity; multidimensional distance; decomposition; combination; prediction; accuracy

I. INTRODUCTION

In computer science field, a recommender system is a relatively new domain of study. Initially, the recommender systems is only a topic of study from several other fields such as cognitive science, approximation theory, information retrieval system, forecasting theory and management science. At the mid of 1990s, the recommender systems become the independent domain of study, i.e. when the researchers have began to focus on the problems of the recommendation using collaborative filtering [1] [2].

The work principle of collaborative filtering algorithm is to generate recommendations for active users based on the opinion history of a group of users that have similarity with that of the active users. The users' opinions are explicitly given in form of rating value [2] [3]. To select new item that will be recommended to the active users, the system previously do the rating predictive value on all of the new items that are not given the rating value yet by the active users. Only the items with highest predictive value will be included into a list of recommendations.

The main problem faced by the collaborative filtering-based recommender system is the prediction accuracy [4]. Many researchers have paid attention to the effort of improving the accuracy, both by developing the prediction technique and the handling of cold-start problem. In this paper, we explain a process of engineering the prediction algorithm on recommender system using multicriteria collaborative filtering to improve the prediction accuracy, including by introducing new approach in user similarity measurement. A metric used to measure the prediction accuracy is Mean Absolute Error defined as : [2][5]

$$MAE = \frac{1}{c}\sum_{i=1}^{c}|r_{ui} - p_{ui}| \qquad (1)$$

where p_{ui} is the user's predictive value u on item i and r_{ui} is a rating value given by the user u on item i, and c is the number of item.

The writing of rest of the paper is arranged systematically as follows. Section 2 provides an explanation of the urgency of multicriteria collaborative filtering in the recommender system. Section 3 explains the process of modifying the suggested multirating prediction algorithm. The testing of prediction accuracy was presented at Section 4, while the discussion of the results of the testing was presented at Section 5. The writing of paper was closed by the conclusion presented at Section 6.

II. THE URGENCY OF MULTICRITERIA COLLABORATIVE FILTERING (MCF) DEVELOPMENT

The collaborative filtering approach is so far largely applied at the recommender systems with only used one criterion to represent the users' opinion on an items [6] [7]. As an example, an individual gives the rating value of 5 in a document, so the value of 5 does not specifically show the criteria of rating used; therefore, a case might occur where

several users give the same values but the criteria used were different. Such problem is called *without distinction of interest problem* [8]–[10].

In order to solve such problems, an idea is offered to accommodate the use of different criteria in making the rating, which is called as multicriteria collaborative filtering [7]. The approach is a variation of the collaborative filtering using many criteria in representing the rating of users' interest. The idea was applied by the Zagat's Guide by determining three criteria of restaurant rating, i.e. *food*, *decor* and *service*, while Buy.com used the multicriteria rating *system* for electronic devices including *display size*, *performance*, *battery life* and *cost*. Yahoo!Movies determined four criteria, i.e. *story*, *action*, *direction* and *visuals* [1].

The use of many criteria in the collaborative filtering is proven to generate recommendation with better quality and more approaching the users' need. The indication of the improving quality can be known from the increasingly high prediction accuracy based on many criteria that are appropriate with the users' tendencies [10] [11]. However, this concept still causes new problems because it is not accompanied by the weighting of criteria reflecting the preferencies of users or frequently called *without weight feature problem* [8]. In order to solve the problem, the weighting is done for several criteria that are regarded as having high priority and the weighting is static in nature. Other criteria regarded as not important were ignored and not involved in the rating determination process.

The static property in the weighting of several criteria and the ignorance of other criteria are potentially harmful to the system, i.e. the lack of prediction accuracy because such users' preferences collectively develop in a dynamic manner. Therefore, it is necessary to develop a multicriteria collaborative filtering that has a capability of improving the weight of criteria adaptively in accordance with the development of the users' collective preferences. The mechanism of updating the weight of criteria should acommodate all the criteria determined, no matter how small the weight of effect on the collaborative process. For the purpose, it is necessary to develop a variation in the multirating value prediction algorithm by combining the concept of classical collaborative filtering and the calculation of criteria weight. The use of many criteria in collaborative filtering also generated an idea to modify a technique for user similarity measurement by the concept of multidimensional distance.

III. Proposed MCF Prediction Algorithm

In the classical collaborative filtering, the model of user profile representation used was the matrix of user-neighborhood where each matrix cell R(u,i) represented the rating value given by user u on item i, with a note that the value 0 indicates the item was never given the users' rating value [12] [13]. The multicriteria collaborative filtering also used the matrix of user-neighborhood to represent user profile, but each user give many ratings to each item, in accordance with the number of criteria determined and added by one overall rating value. Thus, if the number of criteria determined was **k**, each user should give the rating for **k+1**. In the study, the selected object was the scientific documents with four criteria, i.e. topic (k_1), novelty (k_2), recency (k_3) and author (k_4). Thus, the user profile representation also used the matrix of user-items multiratings where each cell of the matrix consisted of five rating values, four for the individual criteria and one for the overall criteria (k_u).

A. User Neighborhood Formation

The formation of user neighborhood is based on user similarity value. The terminology of similarity in this context referred to the similarity in the track records of users in giving ratings on a group of documents. The concept of multicriteria collaborative filtering provides a space for new ideas in calculating the user similarity, i.e. in addition to use cosine, the user similarity can also be measured using the concept of multidimensional distance.

To explain the process of measuring the similarity by using both the models, an example of the matrix of user-document multiratings was given as given in Fig.1 containing eight users, i.e. u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8 and five documents, i.e. d_1, d_2, d_3, d_4, d_5. Each document used four individual criteria and an overall criterion written by k_1, k_2, k_3, k_4, k_u. For example, the users that are active are u_4 and there are three documents that are given the rating value by using u_4, i.e. d_3, d_4 dan d_5. The task of such recommendation system is to make the rating prediction given by u_4 on the three documents, and then give recommendation to the documents with highest predictive value to u_4.In order to do the measurement of user criteria-based similarity, the first step passed through is to decomposite the multicriteria problems become single criterion ones. Results of the decomposition of multiratings given in Fig. 1 become five single criteria matrices as shown in Fig. 2, respectively the matrices for the criteria k_1, k_2, k_3, k_4, and k_u.

User	d_1					d_2					d_3					d_4					d_5				
	k_1	k_2	k_3	k_4	k_u	k_1	k_2	k_3	k_4	k_u	k_1	k_2	k_3	k_4	k_u	k_1	k_2	k_3	k_4	k_u	k_1	k_2	k_3	k_4	k_u
u_1	4	1	1	5	1	3	5	3	2	4	2	3	4	5	4	3	5	4	3	4					
u_2	3	1	5	4	1	2	4	2	1	5						4	3	4	5	4	4	5	4	3	5
u_3	2	3	4	3	3						3	4	4	5	4										
u_4	2	5	4	2	3	5	4	3	3	3															
u_5	4	1	1	3	5	2	3	4	1	1											4	2	4	3	4
u_6	4	4	4	4	4						3	4	5	4	4	5	4	3	4	4					
u_7	5	1	2	4	1	4	5	4	2	4	2	3	4	5	4						4	5	4	5	5
u_8	2	3	4	4	3	4	5	5	4	5															

Fig. 1. User-Document Multiratings

After five matrices were gained, further step was to make the measurement of user similarity for each criteria using cosine formula as follows:

$$s(u,v) = \cos(\vec{R}(u,*), \vec{R}(v,*)) = \frac{\vec{R}(u,*) \bullet \vec{R}(v,*)}{\left\|\vec{R}(u,*)\right\| * \left\|\vec{R}(v,*)\right\|} \quad (2)$$

The algorithm of user similarity measurement using the cosine formula can be written as follows :

```
Input : ratings matrix R(u,i)
Output : similarity(u1,u2,criterion)
1 Set First User and Second User (u1,u2)
```

```
2 For criterion := 1 to 5
3   index := 1
4   For doc := 1 to N
5      if (R(u1,doc) ≠ 0 AND R(u2,doc) ≠ 0)
6      Begin
7       vec_u1[index] := R(u1,doc)
8       vec_u2[index] := R(u2,doc)
9       index := index + 1
10     End Begin
11  End For
12  Sim(u1,u2,criterion):= cos(vec_u1,vec_u2)
13 End For
```

There were five user-neighborhood matrices, so five values of user similarity were obtained, i.e.:

a. $sim_1(u,v)$: user similarity u and v based on topic criteria.
b. $sim_2(u,v)$: user similarity u and v based on novelty criteria.
c. $sim_3(u,v)$: user similarity u and v based on recency criteria.
d. $sim_4(u,v)$: user similarity u and v based on author criteria.
e. $sim_u(u,v)$: user similarity u and v based on overall criteria.

User	k_1				
	d_1	d_2	d_3	d_4	d_5
u_1	4	3	2	3	
u_2	3	2		4	4
u_3	2		3		
u_4	2	5			
u_5	4	2			4
u_6	4		3	5	
u_7	5	4	2		4
u_8	2	4			

(a)

User	k_2				
	d_1	d_2	d_3	d_4	d_5
u_1	1	5	3	5	
u_2	1	4		3	5
u_3	3		4		
u_4	5	4			
u_5	1	3			2
u_6	4		4	4	
u_7	1	5	3		5
u_8	3	5			

(b)

User	k_3				
	d_1	d_2	d_3	d_4	d_5
u_1	1	3	4	4	
u_2	5	2		4	4
u_3	4		4		
u_4	4	3			
u_5	1	4			4
u_6	4		5	3	
u_7	2	4	4		4
u_8	4	5			

(c)

User	k_4				
	d_1	d_2	d_3	d_4	d_5
u_1	5	2	5	3	
u_2	4	1		5	3
u_3	3		5		
u_4	2	3			
u_5	3	1			3
u_6	4		4	4	
u_7	4	2	5		5
u_8	4	4			

(d)

User	k_u				
	d_1	d_2	d_3	d_4	d_5
u_1	1	4	4	4	
u_2	1	5		4	5
u_3	3		4		
u_4	3	3			
u_5	5	1			4
u_6	4		4	4	
u_7	1	4	4		5
u_8	3	5			

(e)

Fig. 2. Results of Decomposition Process

Meanwhile, the measurement of user similarity using the concept of multidimensional distance can be explained in three steps as follows.

The first step is to calculate distance between two users for each document that was *co-rated*. The more the documents that were *co-rated*, the more the values of multidimensional distance. For example, the multiratings of users u were (r_1,r_2,r_3,r_4,r_u) and the multiratings of users were $(r'_1,r'_2,r'_3,r'_4,r'_u)$, so the multidimensional distance between the users u and v for one document was written as d(u,v), calculated by using the Manhattan formula as follows : [14]

$$d(u,v) = |r_1 - r_1'||r_2 - r_2'| + |r_3 - r_3'| + |r_4 - r_4'| + |r_u - r_u'| \quad (3)$$

The second step is to calculate the multidimensional distance between two users based on members D(u, v), i.e. a set of document *co-rated* by the users u and v. The multidimensional distance, written by $d_{total}(u,v)$, was an average of all d(u,v), shown as follows:

$$d_{total}(u,v) = \frac{1}{|D(u,v)|}\sum d(u,v) \quad (4)$$

The third step is to converse the multidimensional distance value gained from the second step to be the similarity value. A relation between multidimensional distance and similarity was stated by [14] with the formula as follows:

$$s(u,v) = \frac{1}{1+d_{total}(u,v)} \quad (5)$$

The algorithm of user similarity measurement by using the concept of multidimensional distance can be written as follows:

```
Input  : ratings matrix R(u,i)
Output : similarity(u1,u2)

1 Set First User and Second User (u1,u2)
2 index := 1
3 For doc := 1 to N
4 if (R(u1,doc) ≠ 0 AND R(u2,doc) ≠ 0)
5     Begin
6       vector_u1[index] := R(u1,doc)
7       vector_u2[index] := R(u2,doc)
8       index := index + 1
9     End Begin
10 End For
11 Distance(u1,u2) := 0
12 For i := 1 to N
13    d_rating[i] := 0
14    For j := 1 to 5
15      d[j] := abs(vector_u1[j]-vector_u2[j])
16      d_rating[i] := d_rating[i] + d[j]
17    End For
18 Distance(u1,u2):= Distance(u1,u2)+d_rating[i]
19 End For
20 Distance(u1,u2) := Distance(u1,u2)/N
21 Similarity(u1,u2) := 1/(1+Distance(u1,u2)
```

B. Prediction Algorithm

The process of the prediction of overall criteria rating can be explained in three steps as follows:

1) Four individual criteria rating prediction

After the database of multiratings was formed, the formation of user-neighborhood and the prediction for the four document criteria, i.e. k_1, k_2, k_3, k_4, can be done using the formula of similarlity-based prediction as follows :

$$P(u,i) = \bar{R}(u,*) + \frac{\sum_{v \in N} s(u,v) * (R(v,i) - \bar{R}(v,*))}{\sum_{v \in N} (|s(u,v)|)} \quad (6)$$

where:

$R(v,i)$: rating value by user *v* on item *i*.
$\bar{R}(u,*)$: average rating value on user *u*.
$s(u,v)$: user similarity *u* and *v*.

Output of the step was four rating predictive values resulted from the system (r'_1, r'_2, r'_3, r'_4) for each document.

2) The calculation of criteria weight

With the step of the prediction of 4 criteria-based individual rating value, the process of computing the relations between the four individual criteria-based rating values (r_1,r_2,r_3,r_4) and the overall criteria saturating value (r_u) was parallelly done based on the multiratings database that was available by using an artificial neural network method. Output of the step was four weights of criteria and one constant e, i.e.:

a. b_1 as the weight for the first criteria (k_1)
b. b_2 as the weight for the second criteria (k_2)
c. b_3 as the weight for the third criteria (k_3)
d. b_4 as the weight for the fourth criteria (k_4)
e. e as computation error (e)

3) The prediction of overall criteria rating

The last step was to predict the rating value for the overall criteria (r'_u) by not using similarity value again, but by utilizing the four individual criteria-based rating value (r'_1, r'_2, r'_3, r'_4) resulted from the first step and four weight of criteria resulted from the second step (b_1, b_2, b_3, b_4) and one constant e, so the overall criteria value was:

$$r'_u = b_1.r'_1 + b_2.r'_2 + b_3.r'_3 + b_4.r'_4 + e. \quad (7)$$

IV. Experiments

To do the testing of the prediction accuracy, some conditions representing the recommender systems was selected, i.e. when the users and document achieved certain amount with certain sparsity level also. The experimental scenario of the testing was as follows:

1) The measurement of Mean Absolute Error (MAE) was done for each criteria.

2) The matrix sparsity level was made various, i.e. 10%, 20%, 30%, 40%, 50% and 60%.

3) The first condition selected for the measurement was when for the first time cold-start problem was solved, where the number of users listed achieved 50 people and the number of document was 100. The second condition was a middle condition, i.e. when the number of users was 100 people and the number of document was 200. In these conditions, there occurred many interraction between users and system where there were the significant addition of new users and documents. The last condition was when the number of user listed achieved 200 people and 400 documents.

4) Prediction rating value for four individual criteria using cosine-based similarity and multidimensional distance-based similarity.

5) Experiment was done for 10 times for each sparsity level.

C. The Prediction Accuracy of Classic Collaborative Filtering

Testing the prediction accuracy of classic collaborative filtering was necessary to do as baseline, with results of the testing shown in Fig.3. From the graphic, it can be seen that the lower the sparsity level of a matrix, the lower the *Mean Absolute Error* (MAE) value. The trend occurred also when the number of users and documents was higher and accompanied by the activity of giving the rating value. The addition of the number of new users and documents into a system but not followed by the activity of giving the rating value will indeed increase a matrix sparsity with impact on the reduced quality of prediction system.

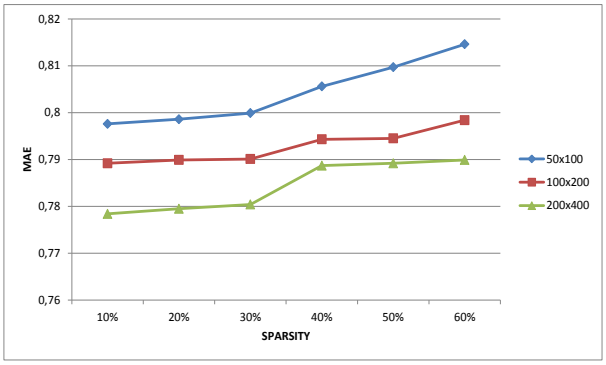


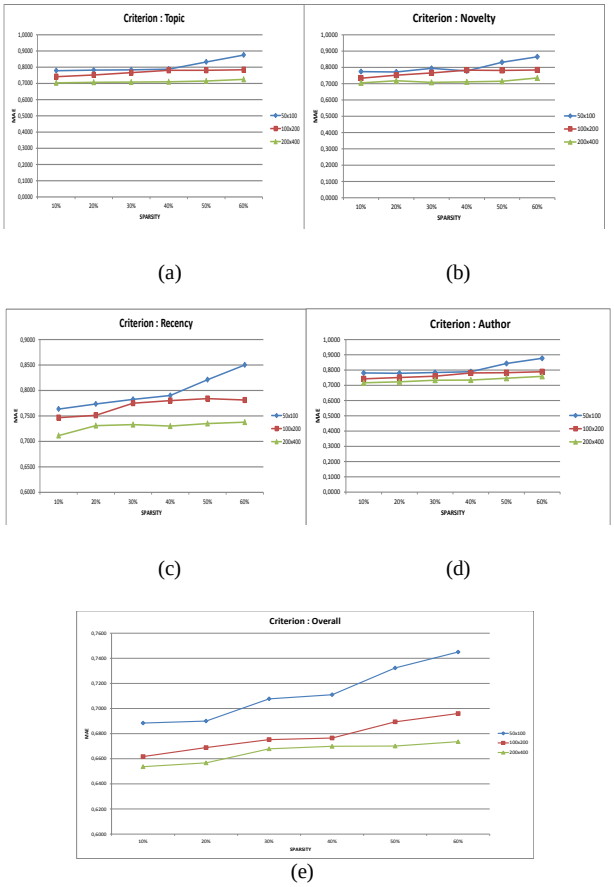
Fig. 3. Graphic of the MAE of Classic Collaborative Filtering

D. The Prediction Accuracy of Multicriteria Collaborative Filtering (MCF) Model.

The measurement of Mean Absolute Error value of the multicriteria collaborative filtering was done in two model variation in accordance with approach used for predicting the four individual rating criteria. The first variation was the model for predicting the four individual criteria rating based on similarity measured using cosine formula as usually done in classic collaborative filtering, while the second variation was the model whose prediction process used the concept of multidimensional distance-based similarity. For the overall rating prediction, both models similarly used a combinatorial technique. Results of the measurement of Mean Absolute Error on the first multicriteria collaborative filtering model was shown in Fig.4.

Results of the measurement of the MAE confirmed the conclusion of previous researchers stating that the more the users actively giving rating value, the more accurate the recommendation produced by collaborative filtering algorithm. In the contrary, although many users listed into a system but when most users will not actively give rating it will indeed weaken collaborative principles as the core of power for recommender systems.

In general, it can be concluded that the best prediction was resulted in the condition of 200x400 with the sparsity level of 10%, while the worst results occured in conditions of 50x100 with the sparsity level of 60%. There was no difference in significant MAE among four document criteria, i.e. topic, novelty, recency, and author. However, the presence of similarity in the trend of MAE value cannot automatically be meant that there were the uniformity of rating value given to four document criteria. It is possible that it was more caused by the users homogenity involved in the system testing process.



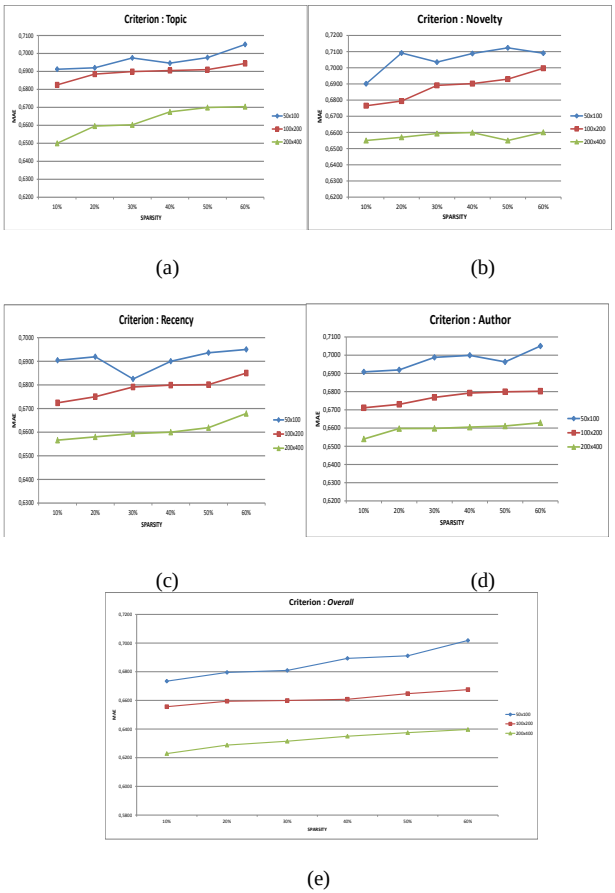
(a) (b) (c) (d) (e)

Fig. 4. Graphic of the MAE of First MCF Model. User Similarity for Each Individual Criteria Was Calculated by Cosine Formula

Results of the measurement of the MAE also provide important information that the overall criteria prediction have better accuracy level compared with four individual criteria, characterized by the lower MAE value for all the sparsity level. Therefore, it can be concluded that the rating value prediction process by using a acombinatorial approach give more accurate results compared with results of pure collaborative filtering approach. In the testing of the model, the lowest MAE value was 0.6537, which was recorded when the number of users and documents was 200x400 with the matrix emptiness level of 10%.

Results of the measurement of MAE for the second model were shown in Fig.5. From the five criteria, there was similarity in trend of predictive values among the four individual document criteria. Meanwhile, for the overall criteria it had better prediction accuracy level. Similar to what happened in the first model, in the second MCF model the best prediction for all the criteria was also resulted in the condition of 200x400 with sparsity level of 10%, while the worst results were also in the condition of 50x100 matrix with the higher sparsity level of 60%.

For individual criteria, the lower value of MAE was 0.6500 that was gained the Topic criteria in conditions of the number of users and documents 200x400 with sparsity level of 10%. Meanwhile, for other three individual criteria, i.e. novelty, recency and author, the lowest value of MAE gained by each was 0.6550, 0.6566 and 0.6540. If compared, the four values of MAE for the four individual criteria did not have significant difference. There were unstable conditions, i.e. when the number of users and documents 50x100 and sparsity level was 20%.



(a) (b) (c) (d) (e)

Fig. 5. Graphic of the MAE of Second MCF Model. User Similarity Was Calculated Using the Multidimensional Distance Approach.

Results of the measurement of the MAE of multi-dimensional distance-based multicriteria collaborative filtering show that rating prediction for the overall criteria has also better accuracy level compared with the prediction for four individual criteria. The best value of MAE for the second model was 0.6229 measured in the conditions of 200x400 with the sparsity level of 10%. The MAE value was lower than the MAE value in the first MCF model, i.e. 0.6537, recorded in the conditions of the same number of users and documents with the same sparsity level. By considering all results of measurement of Mean Absolute Error, it can be concluded that the MCF of second model resulted in more accurate predictive value compared with the MCF of the first model, both for the four individual criteria and the overall criteria.

V. Discussions

Theoretically, the prediction process in a collaborative filtering was actually done based on the principles of similarity value. However, if the number of criteria used is more than one, the overall rating prediction process can be modified by doing a combination between collaborative filtering and criteria weight searching model. However, the way requires conditions, i.e. the availability of user-item ratings database in a large number. By considering all the results of experiment concerning the measurement the Mean Absolute Error shown in Fig.3, Fig.4 and Fig.5 can be known that multicriteria collaborative filtering resulted in the more accurate predictive value than pure collaborative filtering.

The second model resulted in more accurate predictive value compared with the first model, for all individual criteria and overall criteria. It gives new knowledge that although the cosine formula resulted in higher similarity value among users compared with the formula of multidimensional distance, but the prediction accuracy was lower. From computational aspect, overall criteria prediction was more efficient because it only consists of several simple arithmetic statements. However, there were also other computational loads, i.e. when searching the criteria weights using artificial neural network. Periodically, the criteria weights can be updated after there were new rating data.

In addition to give more accurate results of prediction, MCF also given advantage when generating recommendation. It means that some documents that gained high predictive value can be recommended based on the combinatorial criteria. It is very useful for users that want the diversity of recommendation.

VI. Conclusions

Generally, the notion of development of combination prediction algorithm of multicriteria collaborative filtering given the significant increase of prediction accuracy. From the results of experiments, it can be known that average similarity value measured using cosine formula was higher than measured by the concept of multidimensional distance. However, the modification of prediction algorithm using multidimensional distance-based similarity was proven to give more accurate prediction value compared with model using similarity measured by a cosine formula.

References

[1] G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Transaction. Knowledge. Data Engineering*, vol. 17, no. 6, pp. 734–749, Jun. 2005.

[2] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based Collaborative Filtering Recommendation Algorithms," in *Proceedings of the 10th International Conference on World Wide Web*, New York, NY, USA, 2001, pp. 285–295.

[3] M. Ge, C. Delgado-Battenfeld, and D. Jannach, "Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity," in *Proceedings of the Fourth ACM Conference on Recommender Systems*, New York, USA, 2010, pp. 257–260.

[4] A. Umyarov and A. Tuzhilin, "Improving Rating Estimation in Recommender Systems Using Aggregation- and Variance-based Hierarchical Models," in *Proceedings of the Third ACM Conference on Recommender Systems*, New York, NY, USA, 2009, pp. 37–44.

[5] V. Manolis and M. Konstantinos G., "On the Combination of Collaborative and Item-based Filtering," Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece.

[6] P. Resnick and H. R. Varian, "Recommender Systems," *Communication ACM*, vol. 40, no. 3, pp. 56–58, Mar. 1997.

[7] X. Su and T. M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," *Advance Artificial Inteligencel*, vol. 2009, pp. 4:2–4:2, Jan. 2009.

[8] K. Chapphannarungsri and S. Maneeroj, "Combining Multiple Criteria and Multidimension for Movie Recommender System," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Hong Kong, 2009, vol. I.

[9] A. Naak, H. Hage, and E. Aïmeur, "A Multi-criteria Collaborative Filtering Approach for Research Paper Recommendation in Papyres," in *E-Technologies: Innovation in an Open World*, G. Babin, P. Kropf, and M. Weiss, Eds. Springer Berlin Heidelberg, 2009, pp. 25–39.

[10] L. Liu, N. Mehandjiev, and D.-L. Xu, "Multi-criteria Service Recommendation Based on User Criteria Preferences," in *Proceedings of the Fifth ACM Conference on Recommender Systems*, New York, NY, USA, 2011, pp. 77–84.

[11] K. Lakiotaki, S. Tsafarakis, and N. Matsatsinis, "UTA-Rec: A Recommender System Based on Multiple Criteria Analysis," in *Proceedings of the 2008 ACM Conference on Recommender Systems*, New York, NY, USA, 2008, pp. 219–226.

[12] J. A. Konstan, Ed., "Introduction to Recommender Systems: Algorithms and Evaluation," *ACM Transaction Information System*, vol. 22, no. 1, pp. 1–4, Jan. 2004.

[13] M. Montaner, B. López, and J. L. De La Rosa, "A Taxonomy of Recommender Agents on the Internet," *Artificial Intelligence Rev*, vol. 19, no. 4, pp. 285–330, Jun. 2003.

[14] H. Shimodaira, "Similarity and Recommender Systems.", School of Informatics, The University of Eidenburgh, 21-Jan-2014.

Authors Profile

Wiranto is a lecturer at the Bachelor Informatics Program, Faculty of Mathematics and Natural Sciences, Sebelas Maret University in Surakarta. He received both Bachelor and Master of Computer Science from Gadjah Mada University in Yogyakarta, Indonesia. He is currently taking his Doctoral Program at Department of Computer Science and Electronics, Gadjah Mada University.

Edi Winarko is an Associate Profesor and Head of Department of Computer Science and Electronics, Gadjah Mada University in Yogyakarta, Indonesia. He received his bachelor degree in Statistics from Gadjah Mada University, M.Sc in Computer Sciences from Queen's University, Canada and Ph.D in Computer Sciences from Flinders University, Australia. His research interests are data mining, data warehousing and information retrieval. He is a member of ACM and IEEE.

Sri Hartati is an Associate Profesor and Chair of Computer Science Graduate Program, Gadjah Mada University in Yogyakarta, Indonesia. She received her bachelor degree in Electronics and Instrumentation from Gadjah Mada University, both M.Sc and Ph.D in Computer Sciences from New Brunswick, Canada. Her research interests are intelligent systems, decision support systems, medical computing and computational intelligence.

Retantyo Wardoyo is an Associate Profesor and Former Head of Department of Computer Science and Electronics, Gadjah Mada University in Yogyakarta, Indonesia. He received his bachelor degree in Mathematics from Gadjah Mada University, M.Sc and Ph.D in Computer Sciences from University of Manchester, United Kingdom. His research interests are fuzzy systems and expert systems.

Real-Time Simulation and Analysis of the Induction Machine Performances Operating at Flux Constant

Aziz Derouich
Department of Electrical and Computer Engineering
The Higher School of Technology Sidi Mohamed
Ben Abdellah University Fez, Morocco

Ahmed Lagrioui
Department of Electrical and Computer Engineering
The Higher National School of Arts and Trades
Moulay Ismail University Meknes, Morocco

***Abstract*—In this paper, we are interested, in a first time, at the study and the implementation of a V/f control for induction machine in real time. After, We are attached to a comparison of the results by simulation and experiment for, speed responses, flux and currents of the real machine, with a DSPACE card and model established by classical identification (Direct Current test , blocked-rotor test, no-load test , synchronous test), to ensure the validity of the established model. The scalar controlled induction motor allows operation of the motor with the maximum torque by simultaneous action on the frequency and amplitude of the stator voltage, with conservation of the ratio V/f.**

Speed reference imposes a frequency at the inverter supplying the voltages needed to power the motor, which determines the speed of rotation. The maximum torque of the machine is proportional to the square of the supply voltage and inversely proportional to the frequency voltage. So, Keep V/f constant implies a operating with maximum constant torque. The results obtained for the rotor flux and the stator currents are especially satisfactory steady.

Keywords—Induction Machine Modeling; Scalar-controlled induction machine; Experimental identification; Environment Matlab/Simulink/DSPACE

I. INTRODUCTION

The speed control of electrical machine aims to control the operating point of the group "Engine - Load" to best meet the needs of a given industrial application.

This multidisciplinary field currently experiencing considerable importance in industry and in research and requires varied skills in the field of electrical engineering.

Recent progress in the areas of power electronics, automation and digital control led to the development of control system of high performance [1...7].

Today, alternating current machines can replace the direct current machine in most variable speed applications. In particular, induction motor is considered the preferred actuator in constant speed applications. It offers some advantages compared to the DC motor, such as its ease of manufacture and maintenance, without brush-collector device, its weight and low inertia, with an excellent performance. It is also appreciated for its reliability and robustness. However, the simplicity of its mechanical structure is accompanied by a high complexity in the mathematical model (multi-variable and non-linear).

Indeed, in the induction motor, the stator current is used both to generate the flux and the torque. The natural decoupling of the DC machine no longer exists.

On the other hand, we can't know the internal variables of the rotor (rotor current for example) only through the stator. The difficulty is that it exists a complex coupling between the input variables, output variables and the internal variables of the machine, such as torque and speed.

The technological advances had allowed to solve this problem and to develop appropriate strategies of the engine command.

Among them, we can mention the scalar control, Field-oriented control (FOC), direct torque control (DTC), direct mean torque control (DMTC), vector torque control (VTC) and direct self-control (DSC) [8].

The control of AC machine is now almost exclusively based on digital techniques, and so many control algorithms are implemented by the major powers of calculations available. These control algorithms require knowledge of the mathematical model of the machine. The elements constituting of the latter are determined using a set of tests on the machine, especially the Direct Current test, the no-load test, the blocked-rotor test (slip = 1) and the synchronization test (slip = 0). The resulting model is widely used in the stationary regime, that is to say, the machine is assumed to operate at steady regime and powered by a three phase system of constant effective value and rotates at constant speed.

This model is no longer valid if the machine is powered by a three-phase inverter controlled using a control algorithm. This led us to check the validity of this model by comparing the responses in speed, currents and flux of the induction motor obtained by simulation in the Simulink environment with those obtained in real time experimentation.

The scalar control is the easiest control of speed of the induction motor; it allows varying the speed of the machine over a wide range. This is one of the first commands, developed for the variable-speed driving. In this command, we focus on the amplitude of the controlled variable and not to its phase [9]. There are two variants of the scalar control:

- The scalar indirect control where the magnetic flux is controlled by imposing the amplitude / frequency report of the voltage or current

- Direct scalar control where the magnetic flux is controlled from its estimate or its measurement

The second method is more difficult to put into practice. So, we focus in this article on the first approach because it's simple and most used.

II. Modeling The Induction Motor

A good closed-loop control must be based on a mathematical model of the process to be controlled or enslaved. In our application, we use a model of the asynchronous machine that describes the dynamic behavior of the various parameters involved in the control system.

The machine considered in this paper, is a three-phase squirrel-cage (short circuit rotor) asynchronous machine. So, her electrical equations are writing in the following form:

- In the stator : (1)

$$V_{sa} = R_s I_{sa} + \frac{d\emptyset_{sa}}{dt}$$
$$V_{sb} = R_s I_{sb} + \frac{d\emptyset_{sb}}{dt}$$
$$V_{sc} = R_s I_{sc} + \frac{d\emptyset_{sc}}{dt}$$

- In the rotor : (2)

$$V_{ra} = 0 = R_r I_{ra} + \frac{d\emptyset_{ra}}{dt}$$
$$V_{rb} = 0 = R_r I_{rb} + \frac{d\emptyset_{rb}}{dt}$$
$$V_{rc} = 0 = R_r I_{rc} + \frac{d\emptyset_{rc}}{dt}$$

With:

- V_{sa}, V_{sb}, V_{sc} the Three stator voltages.
- I_{sa}, I_{sb}, I_{sc} : the Three stator currents.
- V_{ra}, V_{rb}, V_{rc} the Three rotor voltages.
- I_{ra}, I_{rb}, I_{rc} : the Three rotor currents.
- Φ_{sa}, Φ_{sb}, Φ_{sc} : the flux through the three phases of the stator.
- Φ_{ra}, Φ_{rb}, Φ_{rc} : the flux through the three phases of the rotor.

To replace these differential equations at coefficients which depend on time by simple differential equations with constant coefficients, we apply the Park transformation theory that is the important approach of modeling of induction machine and it's the most used [10].

In our case, we focus on the modeling of induction machine in a reference frame linked to the rotating field. The equations of the machine are then as follows:

- Voltages at the stator : (3)

$$V_{sd} = R_s I_{sd} + \frac{d\emptyset_{sd}}{dt} - \omega_s \emptyset_{sq}$$
$$V_{sq} = R_s I_{sq} + \frac{d\emptyset_{sq}}{dt} - \omega_s \emptyset_{sd}$$

Voltages at the rotor (shorted): (4)

$$V_{rd} = R_r I_{rd} + \frac{d\emptyset_{rd}}{dt} - \omega_{sl} \emptyset_{rq} = 0$$
$$V_{rq} = R_r I_{rq} + \frac{d\emptyset_{rq}}{dt} - \omega_{sl} \emptyset_{rd} = 0$$

- Flux at the stator with Msr = Mrs = M : (5)

$$\emptyset_{sd} = L_s I_{sd} + M.I_{rd}$$
$$\emptyset_{sq} = L_s I_{sq} + M.I_{rq}$$

- Flux at the rotor rotor with Mrs = Msr = M : (6)

$$\emptyset_{rd} = L_r I_{rd} + M.I_{sd}$$
$$\emptyset_{rq} = L_r I_{rq} + M.I_{sq}$$

- Electromagnetic torque: (7)

$$T_e = J\frac{d\theta}{dt} + f.\Omega + T_L$$

With: (d, q) : rotating frame
I_{sd}, I_{sq} : the stator currents in the d-q plane
V_{sd}, V_{sq} : The stator voltages in the d-q plane
I_{rd}, I_{rq} : The rotor currents in the d-q plane
Φ_{sd}, Φ_{sq} : The stator flux in d-q plane
Φ_{rd}, Φ_{rq} : The rotor flux in d-q plane
R_s, R_r : Stator and rotor resistances
L_s, L_r : Stator and rotor Inductances
M : Mutual inductance
ω_s : The stator pulsation
ω : The mechanical pulsation
ω_{sl} : The slip pulsation
J : Moment of inertia
f : Friction coefficient
T_r: Load torque
T_e: Electromagnetic torque
Ω: Mechanical speed ($\omega = p.\Omega$)

III. Principle of Scalar Control

Several scalar controls exist depending on whether it operates on the current or the voltage. They mainly depend on the topology of the actuator used (Udc voltage or Idc current). For our application, we used a voltage inverter supplying the induction machine and drived by the scalar control (ratio V/f constant). The speed variation is achieved by variation of the stator pulsation that is generated directly by the speed controller. This method of control is based on the model of the machine in the stationary regime. For this reason, the study of the machine's model is important in this regime.

The principle of this control is to maintain constant of the ratio V/f, which means keeping the torque constant.

Indeed, if the value of the resistance of the stator windings is neglected, and it is often the case, the electromagnetic torque-slip characteristic in the stationary regime takes the following form:

$$T_e = \frac{3p}{\omega_s}.V_s^2.\frac{\frac{R_r'}{s}}{\left(\frac{R_r'}{s}\right)^2 + (L\omega_s)^2} \quad (8)$$

With
s : Slip

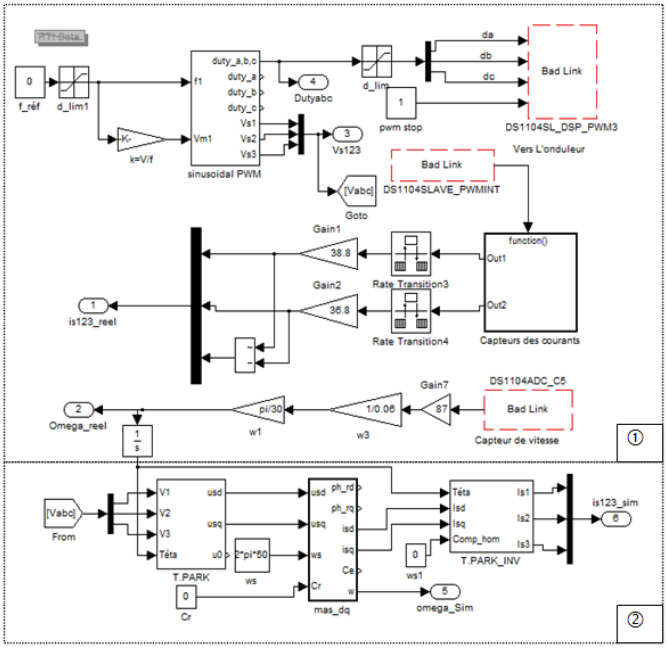
$\frac{R'_r}{s}$: Equivalent resistance rotor conductors reduced to the stator.
p: Number of pole pairs.
ω_s : Stator pulsation.
Vs: Stator voltage.
L: Leakage inductance converted to the primary side.

Often, we are interested in the maximum value of the torque. To calculate it, we look for the value of s that maximizes the expression of the electromagnetic torque Te and then implant it in the previous expression. We give in the equation below the result only:

$$T_{emax} = \frac{3p}{2L}\left(\frac{V_s}{\omega_s}\right)^2 ; \text{ for } s = \frac{R'_r}{L\omega_s} \qquad (9)$$

We deduce that the maximum torque is proportional to the square of the report: $\frac{V_s}{2\pi f_s}$

The command structure allowing the realization of the control at V/f constant of the induction machine is illustrated in the following scheme of Simulink:



①: Experimental part ②: Simulation part

Fig. 1. Simulation scheme

Part ① represents the control law allowing issuing the command signals of the switches (IGBT) of the inverter and the output variables, i.e. the stator currents and rotation speed.

Part ② represents the model of the induction machine in the plane (d-q) with parameters that we want to evaluate, i.e. the stator currents, the rotor flux and the speed of rotation.

IV. Parameters Machine

The induction motor parameters which we have realized our experimentation are shown in the table below:

TABLE I. Parametrs Machine

Parameter	*Value*
Rated power	3 KW
Supply voltage	220V/380V
Synchronous speed	1500 rpm à 50 Hz
Rated speed	1400 rpm
Rated currents (Y/Δ)	7.2A/12.5A
Stator resistance Rs	0.55 Ω
Rotor resistance Rr	0.62 Ω
Pair pole number	2
longitudinal inductance Ld	0.0997 H
Transversal Inductance Lq	0.093 H
Mutual Inductance M	0.093 H
Moment of inertia J	0.01469 kg.m2
Viscous friction coefficient f	0.003035 Nm.sec/Rad

V. Banc Test

The experimental banc Test is consisted of the following elements:

- A cage induction motor having the following characteristics: 3KW, 4 poles, 7.2A/12.5A, 220V/380V, 50Hz, 1400 rpm.
- A diode rectifier providing the DC voltage to the inverter.
- A voltage inverter consists of three bridges, at IGBT and diodes. Three bridges aim to attack the machine and a fourth arm can also be used, when coupled to a resistive load and a suitable command, to protect the electronics of power during braking phases because the diode rectifier is not reversible current and may cause an increase in voltage across the DC bus during braking phases.
- Sensors currents and speed:

- Two current probes LEM HX15-P, LEM P-LV25 for measuring stator currents.
- A tachogenerator 10B0 for measuring the speed.

- The DSPACE-TMS320F240 DSP card ensures the software and digital command aspects. In particular, the digital acquisition of the input signals, the transmission of the inverter bridge control (output signals), current and speed control of the machine.
- The programs, developed under Simulink environment, are implemented within the card. The interface with the operator is then provided by the CONTROL DESK GUI software.
- The DESK CONTROL software is a graphical interface allowing viewing all available variables on patterns Simulink/Dspace. CONTROL DESK combined with DSPACE offers, on Simulink, blocks specific to machine command and allows access to all useful signals to the machine control.

The figure 2 below shows the block diagram of the experimental banc.

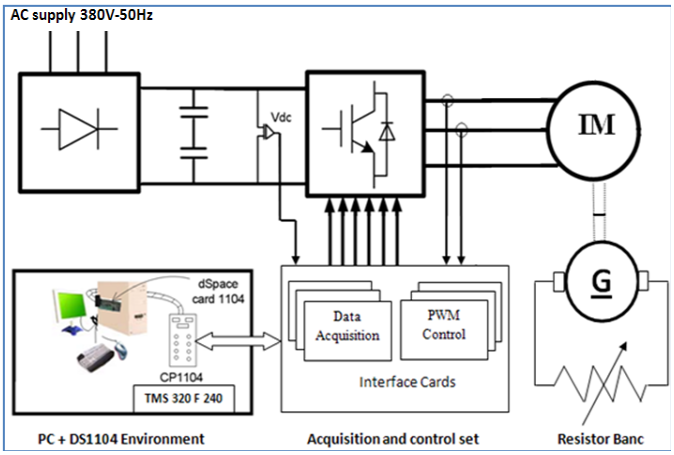


Fig. 2. Block diagram of the experimental banc

The image below shows the real banc experimental [11]:

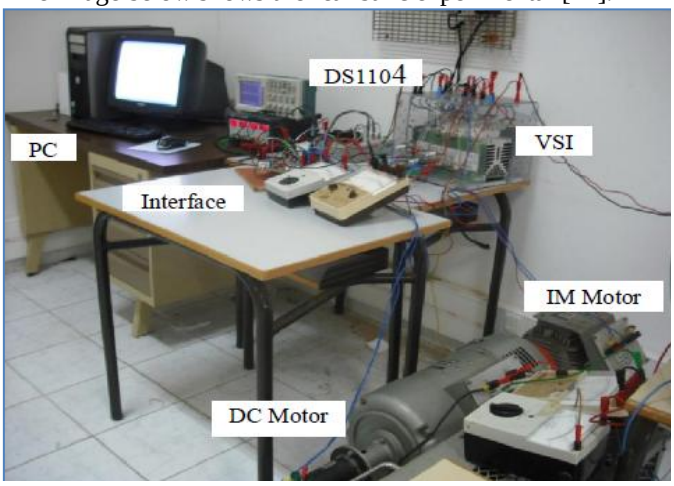


Fig. 3. Experimental banc

VI. Comparison of Experimental Results Versus Those Obtained By Simulation

A. Rotation speed

Response speed is given by the figure 4 below:

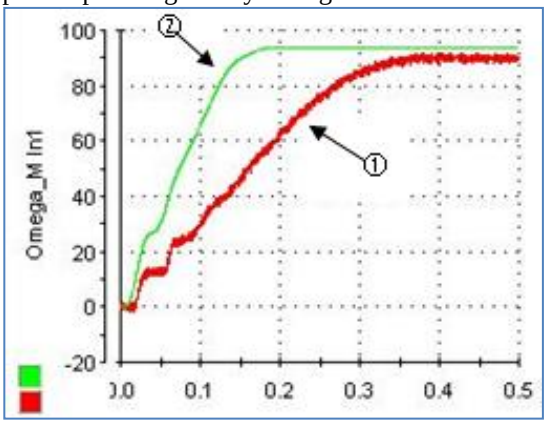


Fig. 4. Rotation speed of the machine

1) *The real speed of the machine.*
2) *Speed of the simulated model.*

We find that:

- The response time of the simulated model is equal to 0.14s
- The response time of the real system is equal to 0.28s.
- The simulated model, as regards the speed, doesn't perfectly follows the machine at the time of starting because all the elements of the machine are still be cold.

B. The stator currents

The stator current of the first phase is shown in the figure below:

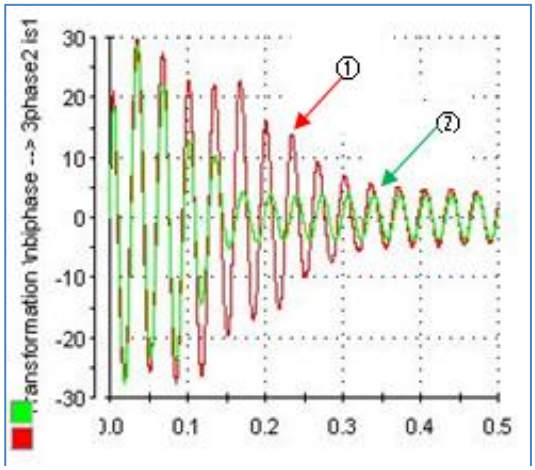


Fig. 5. The stator currents

1) *The real stator current of phase 1 of the machine*
2) *The stator current of phase 1 of the simulated model.*

The figure 6 below shows the various stator currents

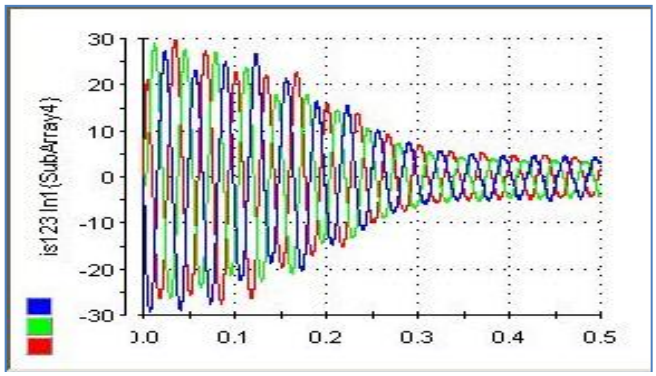


Fig. 6. Stator currents

We note that the current of the simulated model follows perfectly the real current in the stationary regime. It's not the same case in the transient regime.

C. Rotor flux

The shapes of the rotor flux in phase 1 of the machine are shown in Figure 7 below.

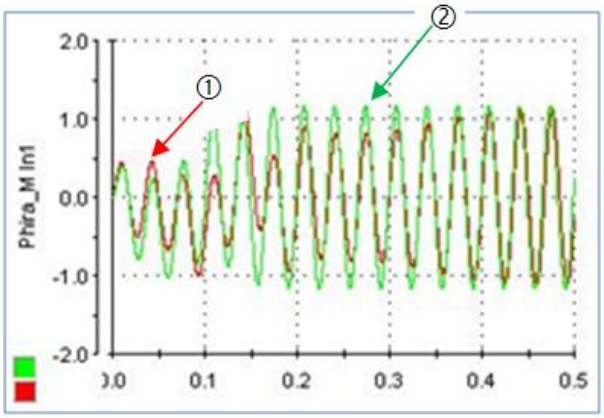


Fig. 7. The rotor flux in phase 1 of the machine

1) The real rotor flux of phase 1 of the machine.
2) The rotor flux of phase 1 of the simulated model.

The shapes of the rotor flux in phase 2 of the machine are shown in Figure 7 below.

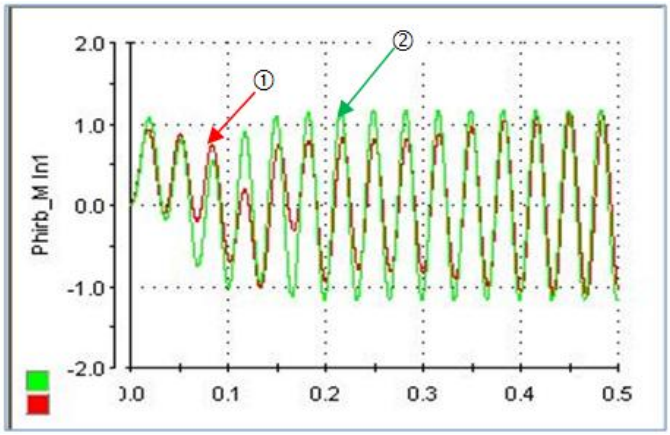


Fig. 8. The rotor flux in phase 2 of the machine

We note, in the figure 7 and the figure 8 above, that the rotor flux of the induction machine in real time and the rotor flux of the equivalent model in the plane of Park converge perfectly in the stationary regime.

VII. CONCLUSION

In this article, we had presented the scalar command of the induction machine operating as a motor. The implementation of the control algorithm explained the performance and limitations of the model of Park established from the values of the parameters of the asynchronous machine obtained by experimental identification (direct current Test, no load, locked-rotor and in synchronism). Indeed, we have seen, from the results, that the model of Park presents some identification errors but which are acceptable and especially in the stationary regime. The errors noticed in the transient regime are mainly due to the moment of inertia of the load on the motor shaft which we have not taken into account.

So an adaptive command that takes into account the parametric variations of the machine (especially the moment of inertia and the load torque) can easily overcome these errors. The perspective of this work consists to implement other algorithms of the induction machine control and compare the result of simulation with the real time.

ACKNOWLEDGMENT

The authors are grateful to the anonymous referees for their valuable comments and suggestions to improve the presentation of this paper. They wish to express their special thanks to all the participants of this work.

REFERENCES

[1] M. Daijyo, I. Hosono, H. Yamada, and Y. Tunehiro, “A method of improving performance characteristics of general purpose inverter,” Trans.Inst. Elect. Eng. Jpn., vol. 109-D, no. 5, pp. 339–346, May 1989

[2] L. Ben-Brahim, “Improvement of the stability of the V/f controlled induction motor drive systems,” in Proc. IIEEE ECON’98, 1998, pp.859–864.

[3] Jaume D, “Commande des systèmes dynamiques par calculateur”, Eyrolles, 91.

[4] Hisham A, Kandil, “La commande optimale des systèmes dynamiques”, Lavoisier, 04.

[5] Pandya, S.N.; Chatterjee, J.K., ”Power Electronics, Drives and Energy Systems (PEDES) & 2010 Power India, 2010- Issue - 20-23 Dec. 2010 pages 1-7.

[6] H.Mahmoudi, A.Lagrioui “Flux-weakening control of permanent magnet synchronous machines”, Journal of Theoretical and Applied Information Technology, 31st December 2011. Vol. 34 No.2 - pp: 110-117.

[7] A.Abbou, H.Mahmoudi, A.Lagrioui, “Comparison of the Rotor Flux Oriented and Direct Torque Flux Control Methods for Induction Motors used in Electric Vehicles”, STA'2006 - Topic ACS "Analysis and automatic Control of Systems ", pp: 1-11.

[8] S ALLIRANI, V JAGANNATHAN, “Direct torque control technique in induction motor drives - a review”, Journal of Theoretical and Applied Information Technology, 28th February 2014. Vol. 60 No.3 – pp: 452-475

[9] R. Toufouti, ”Contribution à la commande directe du couple de la machine asynchrone”, Thèse de Doctorat, Faculté des Sciences de l’Ingénieur, Université Mentouri Constantine, Algérie, 2008.

[10] Sudhir Kumar, P. K. Ghosh, S. Mukherjee ”A Generalized Two Axes Modeling, Order Reduction and Numerical Analysis of Squirrel Cage Induction Machine for Stability Studies”, International Journal of Advanced Computer Science and Applications, Vol. 1, No. 5, November 2010, pp: 63-68.

[11] A. ABBOU, T. NASSER, H. MAHMOUDI, M. AKHERRAZ, A. ESSADKI, ”Induction Motor controls and Implementation using dSPACE”, WSEAS TRANSACTIONS on SYSTEMS and CONTROL, Issue 1, Volume 7, January 2012, pages 26-35.

AUTHORS PROFILE

Mr. Aziz Derouich obtained his diploma from the Superior School of Technical Teaching of Rabat 1995. Further, he got his Diploma of Superior Studies (DESA) in Electronics, Automatic and Information Processing in 2004 and the Ph.D. degree in computer engineering in 2011 from the "University Sidi Mohamed Ben Abdellah" of Fez. He was a professor of Electricity and Computer Science in "Lycée Technique, El Jadida" from 1995 to 1999 and in "Lycée Technique, Fez" from 1999 to 2011.

Since 2011, he is a Professor at the Higher School of Technology, Sidi Mohamed Ben Abdellah University, Fez, Morocco. His research interests

include static converters, electrical machines control, renewable energy and E-learning.

Mr. Ahmed Lagrioui obtained his diploma from the Superior School of Technical Teaching of Rabat 1996. Further, he got his Diploma of Superior Studies (DESA) in 2006 and the Ph.D. degree in electrical engineering in 2011 from Mohammadia School's of engineers, Rabat, Morocco.

He was a professor of Electricity and Computer Science in "Lycée Technique, Taza" from 1996 to 2001 and in "Lycée Technique, Fez" from 2003 to 2014.

Since 2014, he is a Professor at the Higher National School of Arts and Trades, Moulay Ismail University, Meknes, Morocco.

His research interests include static converters, electrical machines control and renewable energy.

On the Performance of the Predicted Energy Efficient Bee-Inspired Routing (PEEBR)

Imane M. A. Fahmy
Computer & Information Sciences department, Institute of Statistical Studies and Research, Cairo University, Giza, Egypt

Laila Nassef
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia

Hesham A. Hefny
Computer & Information Sciences department, Institute of Statistical Studies and Research, Cairo University, Giza, Egypt

***Abstract*—The Predictive Energy Efficient Bee Routing PEEBR is a swarm intelligent reactive routing algorithm inspired from the bees food search behavior. PEEBR aims to optimize path selection in the Mobile Ad-hoc Network MANET based on energy consumption prediction. It uses Artificial Bees Colony ABC Optimization model and two types of bee agents: The scout bee for exploration phase and the forager bee for evaluation and exploitation phases. PEEBR considers the predicted mobile nodes battery residual power and the expected energy consumption for packet reception and relaying of these nodes along each of the potential routing paths between a source and destination pair. In this research paper, the performance of the proposed and improved PEEBR algorithm is evaluated in terms of energy consumption efficiency and throughput compared to two state-of-art ad-hoc routing protocols: The Ad-hoc On-demand Distance Vector AODV and the Destination Sequenced Distance Vector DSDV for various MANET sizes.**

Keywords—PEEBR; Reactive protocol; path selection; MANET; ABC; energy consumption; battery residual power; AODV; DSDV

I. Introduction

The Mobile Ad-hoc Networks MANETs require competent routing protocols since they need to maintain a satisfactory performance as their nodes dynamically move and transmission properties change. Every node in MANETs should achieve two fundamental functions: It acts primarily as a transmitting or receiving point, and as a routing point to relay communicated packets destined for other nodes. Due to the limited communication range of wireless interface, a data packet has to be transferred via several intermediate nodes (Multi-hop routing). Moreover, MANET nodes have limited rechargeable battery power. Thus, the routing mechanism is the most critical and challenging problem in MANETs. In order to solve the routing problem without draining the MANET nodes batteries, a group of MANET energy efficient or power aware routing protocols have emerged as in [1-7].

Swarm Intelligence SI is a computational intelligence approach, as described by [8] that is based on the study of collective behavior of social insects in decentralized, self-organized systems. Ant Colony Optimization introduced by [9] and Bee Colony Optimization by [10] are widely studied among the other Swarm Intelligence techniques applied for networks. Swarm Intelligence SI is a computational intelligence approach, as described by [11]. SI involves a collective behavior of autonomous agents that locally interact with each other in a distributed environment to solve a given problem in the hope of finding a global solution to the problem as defined by [12]. These new SI optimization models have attracted the attention of researchers because they are more robust, reliable, and scalable than other conventional routing algorithms. Since they do not involve more control packets to maintain paths when network topology changes, they are suitable for mobile ad-hoc networks where nodes move dynamically and topology changes frequently. These nature-inspired routing protocols considered the limited resources and highly dynamic environment, as well as the restriction on the exchange of routing information.

Artificial Bee Colony ABC Optimization model proposed in [13] and [14] is a new paradigm of SI that mainly requires two types of agents for routing: scouts, who discover on-demand new routes (paths) to the destinations and foragers, who transport data packets and simultaneously evaluate the quality of the discovered routes based on energy amount expected to be consumed along the path and the nodes batteries residual power. The foragers sense the state of the network, utilize measured metrics to rate different routes in MANET, and then choose the appropriate optimal path for routing of data packets with the aim of maximizing network lifetime.

The Predictive Energy Efficient Bee Routing PEEBR introduced in [16] is a reactive MANET routing algorithm inspired from the natural bees food search behavior. PEEBR's routing technique tends to determine the optimal routing path based on its goodness ratio. The path goodness ratio is a combination of two energetic parameters: the expected energy consumption and the nodes batteries residual power for each potential path.

The paper is organized as follows: The second section presents briefly the routing protocols classification. Some related research works are discussed by the third section. The improved Predictive Energy Efficient Bee Routing (PEEBR) algorithm is described in the fourth section. The improved PEEBR's algorithm simulation results are shown and analyzed in the fifth section. Finally, the sixth section concludes the paper's research goal and future research work.

II. Routing Protocols Classification

In Mobile Ad-hoc Networks MANETs, there are different categories of routing protocols. For unicast routing protocols,

there are four main types of routing protocols according to the routing mechanism employed to discover, control, maintain, memorize or update the path between a specified source and destination nodes in MANET. The proactive routing depends on a routing table stored and regularly updated at each mobile node. While the reactive routing tends to discover a source-destination path on-demand whenever requested. A hybrid routing protocol benefits from both proactive and reactive to make a more reliable and scalable routing by dividing the MANET area into overlapping zones or clusters communicating proactively locally (within the same zone) and reactively to reach a destination in different zone. Finally, in hierarchical routing, each node has a hierarchical ID, which is a sequence of the MAC addresses from the top hierarchy to the source node [15].

According to figure 1, the Destination Sequenced Distance Vector DSDV is a distance vector proactive routing protocol. On the other hand, the Ad-hoc On-demand Distance Vector AODV and the newly proposed Predictive Energy Efficient Bee Routing PEEBR are considered reactive on-demand routing protocols. Finally, the Zone-based Routing Protocol ZRP is a hybrid routing protocol.

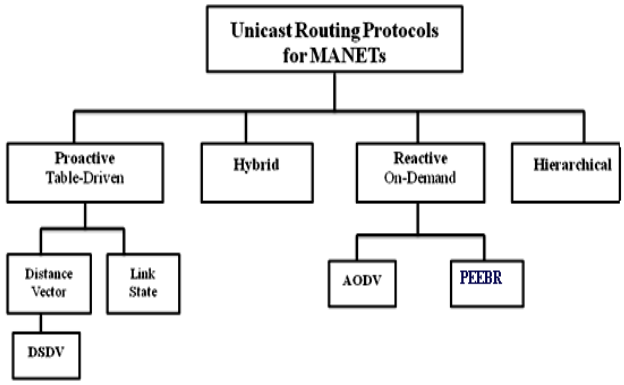


Fig. 1. Unicast MANET Routing Protocols Classification

While AODV, DSDV and ZRP are considered the state-of-art routing protocols for MANETs from the literature, PEEBR and BeeAdHoc [17] could be considered Bio/Nature inspired routing protocols. Among their common features are: multi paths discovery and probabilistic distribution of data traffic on these multi paths to achieve better performance.

The proposed Predictive Energy Efficient Bee-inspired Routing protocol PEEBR by [16] was inspired from the honey bees food search process. Particularly, the two essential groups of bees involved in food source discovery are: The scouts and the foragers. PEEBR inspired by the ABC model is an algorithm for path selection optimization based on energy prediction and consumption efficiency as well as mobile nodes battery residual power maximization in MANETs in order to increase the network lifetime.

III. Related Work

Recently, some research works have emerged for solving the MANET's routing problem and are inspired from the natural bee's behavior as discussed in the following sub-sections.

A. BeeHive Routing Protocol

Wedde, Farooq, and Zhang in [19] introduced a novel routing algorithm called "BeeHive" inspired by the communicative and evaluative methods and procedures of honey bees. In this algorithm, bee agents travel through network regions called foraging zones. On their way, their information on the network state is delivered for updating the local routing tables. BeeHive was fault tolerant, scalable, and relies completely on local, or regional, information, respectively. They have also demonstrated through extensive simulations that the reactive BeeHive routing protocol achieves a similar or better performance compared to state-of-the-art Mobile Ad-hoc Networks routing algorithms such as: AODV, DSDV and DSR.

In BeeHive algorithm, the bee's colony architecture consists of three main exploitation floors as described below:

1) The entrance floor: At this floor the scouts come back to the hive (from their exploration phase). This is the interface to lower level (MAC layer).

2) The dance floor: This is the floor where the dance takes place. The foragers update the routing information of hive's bees (node).

3) The packing floor: This floor is where the worker bees come back with honey to be packed (path control information to update tables). It is responsible of interacting with higher level layer (transport layer).

B. BeeAdHoc Routing Protocol

H. F. Wedde et al. in [17], then presented a new routing algorithm for MANET which is also inspired by the honey bee behavior called BeeAdHoc. The algorithm is simple and mainly needs two types of messages for routing: the scouts: They discover on-demand (reactive) new routes to the destinations. Then, the forgers: which transport data packets and simultaneously evaluate the quality of the discovered routes. The BeeAdHoc routing as shown by figure 2 [17] considers each node in the network as a hive. Each node periodically sends out bee agents: Scouts to explore the network and collect information about any available food sources regardless of their quality. The exploration process achieved by the scout bees could be described and mapped onto the following steps in MANET: Scouts are broadcasted. A TTL (Time To Live packet) is set for each Scout. Then, Scouts take a backward journey to the source (hive) on the same route. At last, Scouts recruit foragers when they are back to the hive by dancing to guide them to the food direction (angle) from the hive.

BeeAdHoc protocol considers the dance floor as the routing table where the bee agents provide the information about the quality of the path they have traversed. Then the exploitation process will be achieved by the foragers and the main workers. Foragers receive data packets from the transport layer (provided by the scouts) and after determining the path quality, they deliver it also by dance to the main workers. Finally, the main workers who receive packets from the transport layer (foragers) are recruited by the foragers such that every worker has a food source.

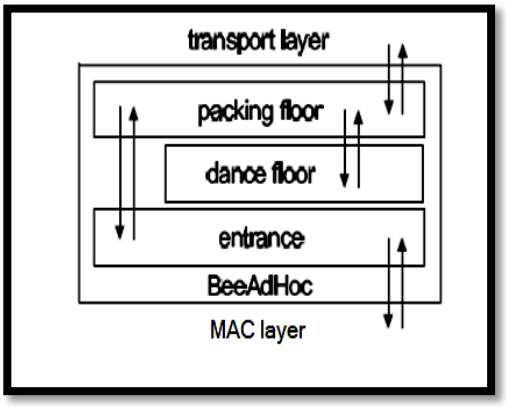


Fig. 2. BeeAdHoc algorithm architecture [17]

It is noteworthy to mention that Beehive and BeeAdHoc routing protocols have not utilized the ABC optimization model in their mechanism or network parameters optimization.

IV. The Improved Predictive Energy Efficient Bee-Inspired Routing (PEEBR)

In bees search process, there are three main phases: First, the scouts seek out all potential food sources which is equivalent to finding all potential MANET routing paths. Then, the foragers assign each discovered food source (routing path) a certain probability according to its quality (nectar amount) interpreted as the link cost for MANET. Finally, the worker bees collect the nectar from the food source with the highest quality according to the qualification probability assigned by the foragers which is equivalent to the optimal path selection according to its quality to communicate the data stream of packets on it in MANETs.

The optimal path selection is based on two main parameters: The average energy consumed by all nodes along each potential path and the nodes average battery residual power together with the hop count. These parameters reflect the path goodness assigned by forager bee agent. The path with the highest goodness ratio should be considered as the optimal path. In PEEBR, the optimal path discovery process from source n_s to the destination node n_d could be described as follows:

A. The Scout Bee

Source node n_s, in order to route efficiently its packets to a destination node n_d, floods a "*Scout packet*" associated with a TTL (Time-To-Live) to all j neighboring nodes. For each "Scout cycle", each "*Scout*" flies over one of the j potential routes R_j until it reaches destination node n_d.

If the TTL packet expires, the "*Scout*" bee agent packet will die indicating failure to reach destination to the source and the corresponding routing path will be avoided.

When a bee agent reaches the destination node n_d, it is sent back to its source n_s through the same traveled route. The backward packet from destination node n_d to source node n_s, "*Scout packet*", collects the potential route's routing information. It counts number of hops $h(R_j)$.

Then, it collects each route nodes residual battery power $B(_{nji})$ where i=1 to N_j nodes and j=1 to M paths. Finally, it memorizes the amount of receiving power consumed.

B. The Forager Bee

At the source node n_s, the *"forager"* evaluation process starts by calculating the predicted amount of energy to be consumed for each "Backward Scout" discovered route. Each potential route cost $f(R_j)$ is calculated for each route Rj dependent on its hop count $h(R_j)$, its nodes residual battery power $B(_{nji})$ and its expected amount of receiving power consumed using expression (6).

The "*Forager*" associates a fitness value $fit(R_j)$ and a goodness ratio of each route $G(R_j)$ as deduced from expressions (7) and (8). At the end of each foraging iteration, each potential path nodes battery residual power $B(n_{ij})$ should be decayed exponentially as computed by (9) to reflect the real world's energy consumption.

Therefore, the optimal route R_o between n_s and n_d is the route with the maximum goodness ratio as given by expression (10).

The other potential routing paths are memorized by source node n_s (for a time interval in communication) in order to be used if any failure occurred during transmission on the optimal route R_o but with respect to their goodness ratio. Finally, a new "Scout cycle" is launched until the maximum number of iterations is reached or a minimal fitness value.

A fault-tolerant and efficient routing protocol is the one that encounters the energy consumption among the other routing information collected before choosing a path and starting transmission. The Artificial Bees Colony ABC model is used by this research in order to employ artificial bee agents that travel from the source node to the destination. The bee agents travel on all potential paths, collect energy information about all the nodes along the path, predict the amount of energy that will be consumed while routing and choose the optimal path. The energy information about a path should reveal:

- ***Each node's battery power residual:*** if it is below a certain predetermined threshold, then the whole path cannot be selected to transmit the data packets
- ***The total energy to be consumed by the path nodes:*** This parameter will indicate the efficiency of the path from energetic point of view, in order to route the packets over the path that consumes less energy. The path that consumes less energy is often with the least number of hops since it will pass by the least number of nodes.

Therefore, the proposed Predictive Energy-Efficient Bee Routing (PEEBR) is assumed to be a reactive routing algorithm that enables a source node to discover the optimal path to a destination node based on the expected energy to be consumed during packets reception and the path nodes residual battery power.

However, PEEBR algorithm could not benefit of the Artificial Bee Colony (ABC)'s food source position optimization functions for the following reasons:

- The nodes random mobility (Any recorded path will not remain the same).
- The reactive nature of the protocol that avoids an inefficient overhead that may be caused by the intention to save and update all paths to all nodes in MANET which results in an inefficient utilization of the MANET's resource: the nodes memory and power.

In table 1, the inspired ABC model's elements are mapped to the PEEBR's algorithm elements together with their optimization interpretation in order to clarify the inspired parts of the ABC model including:

- The fitness function.
- The probability associated with each potential path.

TABLE I. MAPPING ABC MODEL ONTO PEEBR ALGORITHM'S OPTIMIZATION PARAMETERS

ABC	PEEBR	Optimization
Food Source Position	Path between a source node & destination	Possible solution to optimize
Amount of nectar	Average path nodes residual power	Solution quality
Number of employed bees	Number of potential paths	Number of solutions
$f_i = \frac{1}{D_{Train}} \sum_{j=1}^{D_{Train}} d\left(x_j, P_i^{CL_{Known}(x_j)}\right)$	$f(R_j) = h(R_j) \sum_{i=1}^{Nj} \frac{E_r(n_{ij})}{B(n_{ij})}$	Cost Function
$fit_i = \frac{1}{1+f_i}$	$fit(R_j) = \frac{1}{1+f(R_j)}$	Fitness Function
D_{Train}	N_j	Number of nodes along route R_j
SN Sources Number	M paths	Number of Solution
$P_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n}$	$P(R_j) = \frac{fit(R_j)}{\sum_{j=1}^{M} fit(R_j)}$	Probability of solution

The generic expression used to calculate E(p) the energy required to transmit a packet p is given in equations (1) to (5) as in [20]. E(p) in joules (or milli-joules) is given by (1):

$$E(p) = i * v * t_p \quad (1)$$

Where i represents the current consumption, v is the voltage used and t_p is the required time in seconds to transmit a packet given by (2):

$$t_p = \left(\frac{p_h}{6*10^6} + \frac{p_d}{54*10^6}\right) \quad (2)$$

Where p_h is the packet header size and p_d is the packet data size (both in bits). Then, the energy consumed by the node in transmit mode $E_t(p)$ is given by (3), while the energy consumed in reception mode $E_r(p)$ or in overhearing mode when the node overhears the packets exchanged within its range are given by (4):

$$E_t(p) = 280mA * v * t_p \quad (3)$$

$$E_r(p) = E_o(p) = 240mA * v * t_p \quad (4)$$

Therefore, the total amount of energy consumed at a nod n_i is calculated by (5):

$$E(n_i) = E_t(p_{n_i}) + E_r(p_{n_i}) + E_o(p_{n_i}) \quad (5)$$

On the other hand, all nodes residual power $B(n_{ij})$ was initiated using a random value generation in a range from 1000 to 3000 joules. PEEBR's cost function combining the hop count $h(R_j)$ between a given source and destination nodes pair and the average predicted energy consumption $E(n_{ij})$ as path minimizing parameters while the average path nodes battery residual power $B(n_{ij})$ as maximizing parameter are given by (6).

$$f(R_j) = h(R_j) \sum_{i=1}^{N_j} \frac{E(n_{ij})}{B(n_{ij})} \quad (6)$$

Where N_j is the number of nodes on a potential path R_j among M potential paths between the source and destination and the path index j=1,..,M. Then, the path fitness fit (R_j) could be computed using (7).

$$fit(R_j) = \frac{1}{1+f(R_j)} \quad (7)$$

Therefore, the path goodness $G(R_j)$ could be computed using (8)

$$G(R_j) = \frac{fit(R_j)}{\sum_{j=1}^{M} fit(R_j)} \quad (8)$$

In order to test PEEBR's performance, it was run on T_{max}=100 iterations. The nodes battery residual power $B(n_{ij})$ was decayed to reflect the real world's as given by (9):

$$B(n_{ij}) = B(n_{ij})^0 * e^{\frac{-t}{\tau}} \quad (9)$$

Where $B(n_{ij})^0$ is the initial node battery residual power, **t** is the iteration number and **τ** is a time constant. Finally, PEEBR termination conditions were: reaching the maximum number of iterations T_{max} or a minimal predefined fitness value.

The resulting optimal path $\mathbf{R_o}$ is the path with the highest goodness ratio that is given by (10):

$$R_o = \arg\max_j \{G(R_j)\} \quad (10)$$

V. Simulation and Results

In order to evaluate the proposed Predictive Energy Efficient Bee-inspired Routing PEEBR, a self-made simulator using Visual C++ was used to simulate its performance. Since the MANET's critical resource to be efficiently consumed and saved while routing is the nodes battery power, PEEBR's key parameters are: The average energy consumption and the routing path nodes batteries residual power.

In figure 3, it is clearly depicted the impact of increasing the number of nodes in MANET on the the average energy consumed in milli-joules. The proactive DSDV, the reactive AODV and PEEBR protocols showed similar and competitive average energy consumed at smaller MANET sizes as 10 and 20 nodes, then at 30 nodes, their consumption increases more than PEEBR which demonstrates its stability and efficiency. On the other hand, the hybrid ZRP protocol started consuming much less average energy consumption at 10 nodes, then similar to other protocols at 20 nodes, but increased by double at 30 nodes.

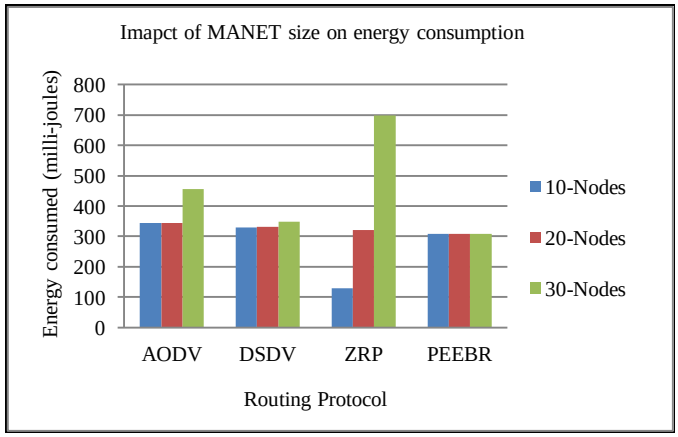


Fig. 3. The impact of varying MANET size on energy consumption by state-of-art routing protocols Vs PEEBR

On the other hand, PEEBR's performance and energy consumption efficiency was compared to another recent bee-inspired routing protocol: BeeAdHoc [17]. Figure 4 shows the impact of increasing the MANET's number of nodes on the energy consumed in transporting one kilobyte of data to its destination which includes the energy consumed for both data and control traffic as defined by [18]. At 10 nodes, PEEBR consumed less energy than BeeAdHoc. Then, at 25 nodes, PEEBR's energy consumption is slightly higher by 0.09 mJ/KB than BeeAdHoc.

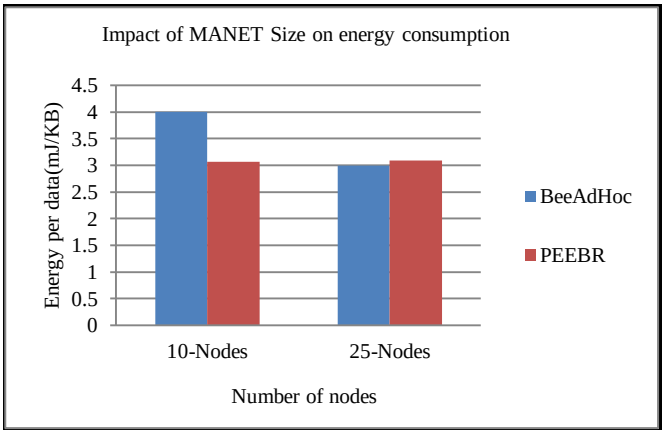


Fig. 4. The impact of varying MANET size on energy consumption by BeeAdHoc Vs PEEBR

VI. Conclusion and Future Works

In this paper, the reactive Predictive Energy Efficient Bee-inspired Routing PEEBR previously proposed in [16] was improved and its path selection optimization algorithm was described. Then, in order to evaluate PEEBR's average energy consumption efficiency in MANETs, we compared its performance to some state-of-art routing protocols as the reactive AODV, the proactive DSDV and the hybrid ZRP. Finally, PEEBR's energy consumed per data measured in mJ/KB was compared to another bee-inspired routing protocol: BeeAdHoc. The simulation results have shown that PEEBR is a competitive energy efficient routing algorithm.

The future work for this research include evaluating PEEBR's performance for other MANET parameters comprising: Packet Delivery Ratio PDR and end-to-end delay under MANET size scenario and mobility scenario.

References

[1] L. M. Feeney. "An energy consumption model for performance analysis of routing protocols for mobile ad hoc networks". Mobile Networks and Applications, 6(3):239–249, 2001.

[2] L.M. Feeney and M. Nilsson. "Investigating the energy consumption of a wireless network interface in an ad hoc networking environment". In Proceedings of IEEE INFOCOM, 2001.

[3] N. Nie and C. Comaniciu, "Energy efficient aodv routing in cdma ad hoc networks using beam forming" EURASIP J. Wireless Communication Networks., vol. 2006, no. 2, pp. 14–14, 2006.

[4] R. Shah and J. Rabaey, "Energy aware routing for low energy ad hoc sensor networks", Wireless Communications and Networking Conference, WCNC2002. IEEE, vol. 1, pp. 350–355 vol.1, 2002.

[5] C. E. Jones, K. M. Sivalingam, P. Agrawal, and J. -C. Chen. "A survey of energy efficient network protocols for wireless networks." Wireless Networks, 7(4):343– 358, 2001.

[6] K. Pappa, A. Athanasopoulos, E. Topalis, and S. Koubias, "Implementation of power aware features in aodv for ad hoc sensor networks a simulation study". IEEE Conference on Emerging Technologies and Factory Automation ETFA, pp.1372–1375, Sept. 2007.

[7] Cui Y., Xue Y., Nahrstedt K., "A Utility-Based Distributed Maximum Lifetime Routing Algorithm for Wireless Networks". Vehicular Technology, IEEE Transactions on vehicular technology, 55(3) (2006) 797, 2006.

[8] Kennedy J, Eberhart R. Particle swarm optimization, In Proceeding of IEEE international conference neural networks, vol. 4; pp. 1942–7, 1995.

[9] M. Dorigo, M. Birattari and Thomas Stutzle. "Ant Colony Optimization: Artificial Ants as computational intelligence technique ». Université libre de Bruxelles, Belgique IEEE Computational Intelligence magazine, November 2006

[10] D. Karaboga and B. Basturk. "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm". Springer Science+Business Media B.V., 2007

[11] Mayur Tokekar and Radhika D. Joshi, "Enhancement of Optimized Linked state routing protocol for energy conservation", CS & IT-CSCP, 2011

[12] J. Wang, E. Osagie, P. Thulasiraman, R. K. Thulasiram, "HOPNET: A Hybrid ant colony OPtimization routing algorithm for Mobile ad hoc NETwork", Elsevier Ad Hoc Networks, June 2008.

[13] D. Karaboga and Ozturk, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm", Elsevier, Applied Soft Computing 11 (2011) 652–657, 2011

[14] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm", Elsevier, Applied Soft Computing 8 (2008) 687–697, 2008

[15] Mehran Abolhasan, Tadeusz Wysocki, Eryk Dutkiewicz, "A review of routing protocols for mobile ad hoc networks", ElSevier Computer Science, Ad Hoc Networks 2 (2004) 1–22

[16] Imane M. A. Fahmy, Laila Nassef and Hesham A. Hefny, "PEEBR: Predictive Energy Efficient Bee Routing Algorithm for Ad-hoc Wireless Mobile Networks", IEEE INFOrmatics and Systems (INFOS2012), 2012

[17] H. F. Wedde, M. Farooq, T. Pannenbaecker, B. Vogel, C. Mueller, J. Meth, and R. Jeruschkat. "BeeAdHoc: an energy efficient routing algorithm for mobile ad-hoc networks inspired by bee behavior." In Proceedings of ACM GECCO, pages 153–160, 2005

[18] Nauman Mazhar, Muddassar Farooq, "Vulnerability Analysis and Security Framework (BeeSec) for Nature Inspired MANET Routing Protocols", GECCO'07, July 7–11, 2007, London, England, United Kingdom, 2007 ACM 978-1-59593-697-4/07/0007

[19] H. F. Wedde, M. Farooq, and Y. Zhang. "Beehive: An efficient fault-tolerant routing algorithm inspired by honey bee behavior". In Proceedings of ANTS Workshop, LNCS 3172, pp. 83–94. Springer Verlag, 2004.

[20] Marco Fotino, Antonio Gozzi, Floriano De Rango, Salvatore Marano, juan-Carlos Cano, Carlos Calafate, Pietro Manzoni, « Evaluating Energy-aware Behavior of Proactive and Reactive Routing Protocols for Mobile Ad Hoc Networks"

Study on Method of Feature Selection in Speech Content Classification

Si An
Institute of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing, China

Xinghua Fan
Institute of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing, China

***Abstract*—Information communication is developing rapidly now, Voice communication from a distance is more and more popular. In order to evaluate and classify the content correctly, the acoustic features is used to analyze first in this paper, Orthogonal experiment[1] method is used to find out characteristic of voice that has contribution to the speech content classification then make it and the textual characteristic together. The result of experiments shows that the feature combination of voice and content has better effect on voice content classification, the effectiveness has been improved.**

Keywords—acoustic features; orthogonal experiment; the SVM classifier; CHI statistical methods; features level fusion; LBS vector quantization algorithm

I. Introduction

As an ideal human-machine communication voice has characteristics of natural, convenient and fast. It has been the pursuit of ideals that making machine understands human speech. In the information age, internet and the telephone exchange increasingly popular, the requirements of the machine is not just to be able to understand the human speech, it also can simultaneously make the appropriate judgment of speech content. For example, to make a fair and impartial judgment of artificial service or sales calls service. Consider this, we began to analysis the voice content, and thus we propose the research of acoustic features. The acoustic features is different, which is caused by two factors, one is the physical structure of the channel itself; the other one is the different vocal habits of everyone, it can also cause the difference of acoustics features that using a different way of vocal organs. So, the voice signals are results of vocal channel structure, pronunciation habits, content of speech and environmental effects comprehensive. It determined by a variety of factors but mainly by semantic [2] content. There are many speech feature parameters existing, but there is no one related only to the voice content or the speaker, We have to choose appropriate parameters to processing or analysis, excluding other influential factors interference caused, and highlight the feature in voice signals which can expression the content to further identify the features that is discriminative to the content category. What we studied is the mostly commonly used parameters of Mel-Frequency Cepstrul Coefficients and time-domain energy and their difference combination.

We can make a judgment and distinguish directly based on text information, but for voice messages, what we first should do is to do speech recognition, in this process, we have a few things to do, such as pre-process, feature selection, the structure of acoustics model and language model. After these procedure, the received content is different from the initial meanings which will cause the miss of the result of classification. So in order to get precisely analysis and estimate, it's necessary to composite the two kind of information. That is data fusion. We can train the classifier by using the feature vectors that combining the acoustics features and text feature. But because these two features come from different time domain and they have distinct criterion, the dimension of the space will increase if merging them simply. So we'd better find the optimum combination of features to improve the robustness of the system.

II. The Feature Analysis of Speech

A. Acoustic feature extraction

At present, the researchers find the characteristics that are closely related to the pronunciation are mainly pitch frequency, short-time energy. The parameter of formant and the spectral energy distribution is related to the sound way. In this paper, What we studied is the commonly used parameters of Mel-Frequency Cepstrum Coefficients and time-domain energy and their difference combination. Because the first-order differential cepstrum parameters reflected the changes over time that its dynamic characteristics, we think it can complete more express the original speech.

1) The optimization of the acoustic features

When analyzing the content of speech, it has to have strong interference ability of environmental noise and robustness. It can't meet the requirements of robustness if sticking with single parameters. Selecting the number M from the given number of N feature parameters X(1),X(2),...X(N) to training the classifier, it's called feature selection. There are some ways of feature selection in other field of research, in paper [3], there is a method named multi-objective optimization. If only these features are put together freely, the dimension of features will be quite high which not improve the performance of the system but extend the training time thus affect the real-time performance. It is not convenient to use. How to get the information that has the characteristics of the complementary role from the large number of feature parameters, it's a problem with practical significance. In the following, we analyze and optimize the characteristic parameters by orthogonal experiment.

2) Orthogonal experimental design

Orthogonal experimental design is widely applied in agriculture, process design in the developed country. There are many examples of successful application in our country[4,5].The method of orthogonal experimental design has made good use of the table—"Orthogonal" to arrange experiment.

It can elected strong small number of experimental conditions in many experiments and inference to find the best process conditions through these number of experimental conditions[6].The factors called factor which can affect the result of the experiment in this way, the state of different factor called level. Orthogonal experiment is to find the optimum combination of the factors exists. In the process of searching the required test times is fewer than in the exhaustive method. For instance, in the experiment of this paper, there need two to the power of twenty six in exhaustive method, but it just need thirty two in orthogonal experiment.

3) Orthogonal experiment steps

a) Constructing orthogonal table: Orthogonal table is usually need solid mathematical theory, but when the factor level is two, the table is very easy to construct, reference the literature [7].

*b) Factors:*In the orthogonal table, the amount of each level is equal to the average and between any two columns of different levels of the total number of combinations is equal to the average, so, when arranging orthogonal experiments, all sorts of factors collocation is balanced. In the table, every row said an experiment scheme, which is a combination of various factors in state; each column figures show that the corresponding factors of the state.

c) The experiment results analysis:Analysis of variance is that can distinguish the difference between experimental results and error caused by the fluctuation of differences between the experimental results, this method of math make up the deficiency of the poor analysis method in this respect, so, in this paper, we using the analysis of variance to experimental results. According to the theory of difference analysis, we can get the discriminative that the change of level caused by the difference between the experimental results. If the experimental results changed caused by the changes of factor levels within the error range or has little difference with the error, the change of this factor level can determined not cause a significant change in results; On the other hand, if factor levels' change will not cause changes than error range in the experimental results, we can sure that the factor has a significant impact on the experimental results. The purpose of this analysis is to find out the things which have a significant impact factor through the data.

*d) The selection of orthogonal table and the structure:*For 2 levels orthogonal table, Hadamard horse matrix can be used to construct: let the second-order matrix

Hardamad is the basic matrix, the rest can be done in the same manner, in this paper we use the table L_{26} (2^{26}) .Remove the first column which is full one, for simplicity, remove the back of the five columns and turn -1 to 0, the result was a matrix of 32*26.

B. Speech content textual feature extraction

In addition to the acoustic features, the textual feature should also be considered. There is some correlation between the audio data and content data, the incomplete of the audio can be added in some ways for example the text information. So the most direct way to evaluate the speech is do classification by using the text after recognizing.

SVM is a method of sample learning that has a solid theoretical foundation. It implements an efficient "transduction reasoning" from training sample to predict samples and simplifying the classification. With the support vector machine classification's better overall performance, it is used in this paper. For content, the word frequency is the characteristic of the text.

1) The method of textual feature selection

There are two factors can be observed in the text in fact, that is word frequency and document frequency, there are some feature selection algorithm based on the document frequency such as CHI statistics, Information Gain(IG), Mutual Information(MI). Many experiments show that the CHI statistics is more commonly used method. Its basic idea is to determine the theory correct by observing the deviation of the actual value and the theoretical value or not. It is a measure of the relevance between feature word t and document category C_j, assuming that meet the distribution of the first order χ^2 between t and C_j . The bigger of the value of chi-square statistic the key words belong to a category the greater the relevance between the key words and the category. The Chi-square statistic calculated which the key words t to the category C_j is defined as :

$$\chi^2(t, c_j) = \frac{N \times (AD - BC)}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (2)$$

In the formula (2), A stands for the number of documents which belong to class C_j and contains the key word t; B stands for the number of documents which not belong to class C_j but contains the key word t; C stands for the number of documents which belong to class C_j but not contains the key word t; D stands for the number of documents which neither belong to class C_j nor contains the key word t; N represents the total number of total text in training corpus. In this paper, we simply believe that the characteristic words are the words which has high CHI value.

III. The Structure of The Experiment

Audio content classification [7] is that to train a classifier by using the extracted features data. In order to train the classifier, we should build a data set used in the experiments, which includes voice files and the text files after recognizing. SVM is used to training the data set to get the classifier, the average value of many experiment is taken as the final result. The experiment is divided into four parts in this paper: the first part is training the audio features individually; the second part is

doing orthogonal experiment to get the optimized combination; the third part is training the classifier by using the text features (we can considered it as ideal that the text file is not the result of speech recognizing); The forth is training the classifier by the fusion features combined the acoustic features with textual features.

The overall structure of experiment in this paper is showed in figure1.

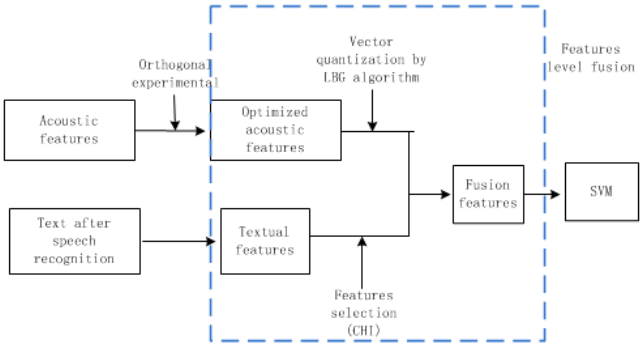


Fig. 1. The overall structure in this paper

A. Design of experiment and result analysis

1) Corpus sources : The source of experiment data in this experiment is: the real situations of dialogue from two men and three women who are articulate simulate the real scene. There are 2179 dialogues, the content of dialogue are two parts for transportation and legal launched two topics. There are 1045 speech data about traffic class, 1134 speech data about legal. The sampling frequency is 16 KHz, The quantitative accuracy is 16bit, and the voice of the frame length is 256 sampling points. In these experiments, because the signal noise ratio of data is low, we have something to do to improve the efficiency of endpoint detection; according the paper [9], when to set the threshold to time we set a new threshold by weighting the maximum and minimum values of the volume and averaged it. More accurate effective interception of voice is got and it provides favorable conditions to feature selection.

2) Experimental evaluation criteria: The following indicator is used to evaluate the performance of the classification results of the experiment,

Precision P that is also named accuaracy*:*

$$p = \frac{A}{B} \times 100\% \quad (3)$$

Recall R:

$$R = \frac{C}{D} \times 100\% \quad (4)$$

F_1 value :

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\% \quad (5)$$

In the formula (3), A stands for the number of text that determined correctly by classifier, B represents the number of text determined by classifier.

In the formula (4), C stands for the number of text that determined correctly by classifier; D represents the number of text in the test set. The F_1 value is a comprehensive evaluation standard.

B. Experiment

1) Using the audio features only to train the classifier:

The classifier is trained by the audio features only, the test result is:

TABLE I. ACCURACY OF ORIGINAL ACOUSTIC FEATURES

Features	Dimensions	accuracy%	F_1 %
Acoustic	26	74.9	70.3

2) Optimize the audio feature parameters by orthogonal experiment

*a) Factor selection:*The selection of the characteristic parameters of a total is 26 in this paper, each factor has two levels which 1 stand for used and 0 is unused. Characteristic parameters including the combination of MFCC feature and energy mentioned above and their dynamic first-order difference, a total of 26 dimension:

$mfcc_1, mfcc_2, mfcc_3, mfcc_4, mfcc_5, mfcc_6, mfcc_7, mfcc_8, mfcc_9,$

$mfcc_{10}, mfcc_{11}, mfcc_{12}, En, mfcc_{13}, mfcc_{14}, mfcc_{15}, mfcc_{16}, mfcc_{17},$

$mfcc_{18}, mfcc_{19}, mfcc_{20}, mfcc_{21}, mfcc_{22}, mfcc_{23}, mfcc_{24}, \Delta En$

Because the first-order differential cepstrum parameters reflected the changes over time that its dynamic characteristics, we think the dynamic characteristic parameters can complete more express the original voice.

b) The experiment design

According to the design of orthogonal experimental design method, each column corresponds to characteristic parameters and each row represents a kind of combination plan, the number 1 stands for used and 0 is unused. The last column of the table is each set of features combination experiment by the end of the audio classification effect. The experimental scheme is shown in table 2.

After orthogonal experiment result is analyzed from the table, the P value in the statistical is obtained by look-up table after getting F value, every parameter was coded B1,B2,B3...:

It can be seen through the analysis of significance that the audio content for the classification result is greatly influenced by B3,B5,B6,B8,B11,B14,B17,B18,B25.

That is the parameters:

$mfcc_3, mfcc_5, mfcc_6, mfcc_8, mfcc_{11}, mfcc_{14},$

$mfcc_{17}, mfcc_{18}, mfcc_{25}$

the parameters has little effect on experimental results are:

$mfcc_9, mfcc_{13}, mfcc_{21}, mfcc_{22}, mfcc_{23}$

TABLE II. THE ACCURACY OF DIFFERENT FEATURE COMBINATION

No.	Combined Solutions	accuracy %
1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	74.9
2	0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0	66.5
3	1 0 01 1 0 0 1 1 0 0 1 0 1 0 1 1 0 0 1 1 0 0 1 1 0	70
4	0 0 1 1 0 0 1 1 0 0 1 1 0 1 1 1 0 0 1 1 0 0 1 1 0 0	64
5	1 1 1 0 0 0 0 1 1 1 1 0 0 1 1 0 0 0 0 1 1 1 1 0 0 0	68
6	0 1 0 0 1 0 1 1 0 1 0 0 1 1 0 0 1 0 1 1 0 1 0 0 1 1	69
7	1 0 0 0 0 1 1 1 1 0 0 0 1 1 0 0 0 1 1 1 1 0 0 1 0 1	65
8	0 0 1 0 1 1 0 1 0 0 1 0 0 1 1 0 1 1 0 1 0 0 1 0 1 0	74.5
9	1 1 1 1 1 1 1 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0	69
10	0 1 0 1 0 1 0 1 1 0 1 0 1 1 0 1 0 1 0 0 1 0 1 0 1 0	64
11	1 0 0 1 1 0 0 1 0 1 1 0 1 1 0 1 1 0 0 0 0 1 1 0 0 1	68
12	0 0 1 1 0 0 1 0 1 1 0 0 0 1 1 1 0 0 1 0 1 1 0 0 1 1	68.5
13	1 1 1 0 0 0 0 1 0 0 0 1 1 1 1 0 0 0 1 0 0 1 0 1 1 1	73.5
14	0 1 0 0 1 0 1 0 1 0 1 1 0 1 0 0 1 0 1 0 1 0 1 1 0 1	71
15	1 0 0 0 0 1 1 0 0 1 1 1 0 1 0 0 0 1 1 0 0 1 1 1 1 0	72.5
16	0 0 1 0 1 1 0 0 1 1 0 1 1 1 1 0 1 1 0 0 1 1 0 1 0 0	69
17	1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	73.8
18	0 1 0 1 0 1 0 1 0 1 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0	71.5
19	1 0 0 1 1 0 0 1 1 0 0 1 0 0 1 0 0 1 1 0 0 1 1 1 0 1	68.5
20	0 0 1 1 0 0 1 1 0 0 1 1 1 0 0 0 1 1 0 0 1 1 0 0 1 1	69.5
21	1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1	70.5
22	0 1 0 0 1 0 1 1 0 1 0 0 1 0 1 1 0 1 0 0 1 0 1 1 0 1	67.5
23	1 0 0 0 0 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 1 1 1 1 0	73.5
24	0 0 1 0 1 1 0 1 0 1 1 0 1 1 0 1 0 0 1 0 1 1 0 1 0 0	71.5
25	1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1	73.5
26	0 1 0 1 0 1 0 0 1 0 1 0 1 0 1 0 1 1 0 1 0 1 0 1 0 1	69.5
27	1 0 0 1 1 0 0 0 0 1 1 0 1 0 1 0 0 1 1 1 1 0 0 1 1 0	70
28	0 0 1 1 0 1 1 0 1 1 0 0 0 0 1 0 1 1 1 0 1 0 0 1 1 0 0	69.5
29	1 1 1 0 0 0 0 0 0 0 0 0 1 1 0 0 1 1 1 1 1 1 1 1 0 0 0	69.1
30	0 1 0 0 1 0 1 0 1 0 1 1 0 0 1 1 0 1 0 1 0 1 0 1 1 0	70.2
31	1 0 0 0 0 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 0 0 1 0 1	70.6
32	0 0 1 0 1 1 0 0 1 1 0 1 1 0 0 1 0 0 1 1 0 0 1 0 1 1	72.4

We got the characteristics of the combination that have a greater impact on experiment in theory, but whether it is effective in practice or not, in order to test whether the result is optimal, put these features into the experimental group and then trained classifier.

Finally obtained as shown in table:

TABLE III. THE CONTRAST TABLE

	dimensions	accuracy%	F_1 %
The original feature parameters	26	74.90	70.3
The combined scheme	13	79.00	77.5

It can be seen that when do experiment with 26 d characteristic at orthogonal experiment, the classification accuracy is 74.9%. After the theoretical analysis, the most influential characteristic parameters for the efficiency of experiments is found, taking them into the subsequent experiments, the accuracy is 79 %, it has been improved and the feature space is reduced and the F_1 is also improved from 70.3 % to 77.5%.

C. Training the classifier using textual features

When analyzing the audio text after speech recognizing, features selection is the first step. SVM is used to training the classifier by using the CHI statistic value, then the classifier is on the test set for testing.

TABLE IV. THE ACCURACY OF TWO KIND OF TEXT

Textual features(CHI)	Original text	Recognized text
accuracy %	90.15	84.20
F_1 %	88.60	82.35

D. Features fusion

In the study of audio classification, there are two ways usually: features fusion and decisions level fusion [8]. Features fusion means that extracting features from the audio files and audio text respectively, then training the classifier by the merged features. Decisions level fusion is that training classifier by the acoustic features and text data individually, then taken together the results in some way. In this paper, we used the method of features fusion. The biggest problem in features fusion is that the level of the phonetic characteristics and the text characteristics. For example, it's hard to say the relationship between the energy and the classes. So, the primary problem in features fusion is the conversion that the characteristics of the two form to a level. For the extracted audio features, they should be mapped to the text like "audio word". Quantization algorithm is used to achieve this mapping. In order to reduce the complexity of the calculation, The algorithm of LBG-VQ[9] is used to get the codebook from the training data. Once acquired the codebook, the feature vectors are mapped to the nearest "audio word" based on the codebook. After getting the "audio word", the TF-IDF[101 weighting function is used to calculate the weight that the "audio word" in the "audio text". Under the assumption that these two forms of audio and text "key words" were independent of each other. So we spliced together them directly to complete the fusion of these two kinds of pattern characteristics. The experimental results as shown in the figure below:

TABLE V. THE RESULT CONTRAST

Features	accuracy %	F_1 %
A（Textual features）	84.20	82.35
A+B(Optimized acoustic + textual features)	87.30	85.45
C+B(Optimized Acoustic + textual features)	92.25	90.40

We can see that the accuracy of classification is improved after features fusion. The fusion of the phonetic features optimized and the textual features lead to the improved accuracy of the classifier, the accuaracy is improved from 84.2% to 92.25% andt the the F_1 value is improved from 82.35% to 90.40%.So it can be concluded that combining characteristics effectively can training better classifier.

IV. CONCLUSION

When analyzing the speech, we usually separate the acoustic part and its semantic part to deal, it will lead to lose their complementary part and cause mistakes. In this paper, the acoustic features are optimized firstly and combined with the semantic characteristics then, the classification results were improved and it cost less time. It proves that the method of features fusion is effective. In the process of the fusion of the two kinds of features, we string them together simply, not considering the distribution of their weights; it's what we should do next.

REFERENCES

[1] YANG Da-Li, XU Ming-Xing, WU Wen-Hu.Study of Feature Selection for Speech Rccognition[J]. JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT. 2003, 40(7): 963-939.

[2] SHAN Song-wei, FENG Shi-cong, LI Xiao-ming. A comparative study on several typical feature selection methods for Chinese web page categorization[J]. Computer Engineering and Applications, 2003(22):146-148.

[3] E. Zitzler and L. Thiele. (1999) "Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach."IEEE Transactions on Evalutionary Computation, vol. 3. pp. 257-71.

[4] Chunyuan, Zhao "Discussion on the application of multimedia teaching on classroom teaching of advanced mathematics"[J]. Journal of shenyang Institute of Engineering(Social Sciences). Jul. 2010, Vol. 6, No.3, pp. 399-400.

[5] Ho S Y, Lin H, Li W H, et al. Orthogonal particles swarm Optimization and its application to task assignment problems [J].IEEE Transactions on Systems, Man and Cybernetics, 2008,38(2) :288－298

[6] Li X L, Zhao Q, Zhang C J. Research on multiple index optimization method of the orthogonal test design[C] // IEEE International Conference on Computer Science and Information Technology.[s.l.]:[s.n.], 2010:224－226.

[7] Bai Liang; Hu Yaali; Lao Song yang; Chen Jianyun; Wu Lingda; Feature analysis and extraction for audio automatic classification. IEEE International Conference on Volume 1, 10-12 Oct.2005:767-772

[8] Z. Zeng, Y. Hu, M. Liu, Y. Fu, and T.S. Huang, Training Combination Strategy of Multi-stream Fused Hidden Markov Model for Audio-visual Affect Recognition[C].Proc.14th ACM Int'l Conf.Multimedia(Multimedia'06),2006:65-68

[9] A.Gersho and R.M.Gray.Vector quantization and signal compression [M].Norwell, MA, USA: Kluwer Academic Publishers, 1991.

[10] Kazunari Sugiyama et al. Refinement of TF-IDF Schemes for web Pages using their Hyperlinked Neighboring Pages, 14th ACM Conference on Hypertext and Hypermedia 2003, Pages 198-207

Comparative Performance Analysis of Wireless Communication Protocols for Intelligent Sensors and Their Applications

Chakkor Saad
Department of Physics,
Team: Communication and Detection Systems,
University of Abdelmalek Essaâdi, Faculty of Sciences,
Tetouan, Morocco

Baghouri Mostafa
Department of Physics,
Team: Communication and Detection Systems,
University of Abdelmalek Essaâdi, Faculty of Sciences,
Tetouan, Morocco

El Ahmadi Cheikh
Department of Physics,
Team: Communication and Detection Systems,
University of Abdelmalek Essaâdi, Faculty of Sciences,
Tetouan, Morocco

Hajraoui Abderrahmane
Department of Physics,
Team: Communication and Detection Systems,
University of Abdelmalek Essaâdi, Faculty of Sciences,
Tetouan, Morocco

***Abstract*—The systems based on intelligent sensors are currently expanding, due to theirs functions and theirs performances of intelligence: transmitting and receiving data in real-time, computation and processing algorithms, metrology remote, diagnostics, automation and storage measurements…The radio frequency wireless communication with its multitude offers a better solution for data traffic in this kind of systems. The mains objectives of this paper is to present a solution of the problem related to the selection criteria of a better wireless communication technology face up to the constraints imposed by the intended application and the evaluation of its key features. The comparison between the different wireless technologies (Wi-Fi, Wi-Max, UWB, Bluetooth, ZigBee, ZigBeeIP, GSM/GPRS) focuses on their performance which depends on the areas of utilization. Furthermore, it shows the limits of their characteristics. Study findings can be used by the developers/ engineers to deduce the optimal mode to integrate and to operate a system that guarantees quality of communication, minimizing energy consumption, reducing the implementation cost and avoiding time constraints.**

Keywords—Wireless communications; Performances; Energy; Protocols; Intelligent sensors; Applications

I. INTRODUCTION

Wireless technologies have made significant progress in recent years, allowing many applications in addition to traditional voice communications and the transmission of high-speed data with sophisticated mobile devices and smart objects. In fact, they also changed the field of metrology especially the sensor networks and the smart sensors. The establishment of an intelligent sensor system requires the insertion of wireless communication which has changed the world of telecommunications. It can be used in many situations where mobility is essential and the wires are not practical.

Today, the emergence of radio frequency wireless technologies suggests that the expensive wiring can be reduced or eliminated. Various technologies have emerged providing communication differently. This difference lies in the quality of service and in some constraints related on the application and it environment. The main constraints to be overcome in choosing a wireless technology revolve around the following conditions [1], [2]:

- Range
- Reliability
- Bandwidth
- conformity (standards)
- Security
- Cost
- Energy consumption
- Speed and transmission type (synchronous, asynchronous)
- Network architecture (topology)
- Environment (noise, obstacles, weather, hypsometry)

In this work, we studies using a comparative analysis, the different parameters which influence the performance and quality of a wireless communication system based on intelligent sensors taking into our consideration the cost and the application requirements.

We can classify the requirements of applications using smart sensors into three main categories as shown in table I.

TABLE I. NEEDS BASED APPLICATIONS

Types of application	Specifications and Needs
Environmental monitoring	▪ Measurement and regular sending ▪ Few data ▪ Long battery life ▪ Permanent connection
Event detection	▪ Alerte message ▪ Priority ▪ Confirmation statuts ▪ Few data ▪ Permanent connection
Tracking	▪ Mobility ▪ Few data ▪ Localization ▪ Permanent connection

II. RELATED WORK

In the related work, many research studies in [3-8] have been focused on wireless sensor networks to improve communication protocols in order to solve the energy constraint, to increase the level of security and precision and to expand autonomy for accuracy, feasibility and profitability reasons. On the other side, the field of intelligent sensors remains fertile and opens its doors to research and innovation, it is a true technological challenge in so far as the topology and the infrastructure of the systems based on intelligent sensors are greatly different compared to wireless sensor networks, particularly in terms of size (number of nodes) and routing. In fact, to preserve the quality of these networks, it is very difficult even inconceivable to replace regularly the faulty nodes, which would result in a high cost of maintenance. The concept of energy efficiency appears therefore in communication protocols, [5-9]. Thus, it is very useful to search the optimization of data routing and to limit unnecessary data sending and the collisions [6], [9]. The aim challenge for intelligent sensors systems is to overcome the physical limitations in data traffic such as system noise, signal attenuation, response dynamics, power consumption, and effective conversion rates etc… This paper emphasis on the metrics of performance for wireless protocols which stands for superior measurement, more accuracy and reliability. The object of this study is for realizing an advanced intelligent sensors strategy that offers many system engineering and operational advantages which can offer cost-effective solutions for an application.

III. NEW CONSTRAINTS OF INTELLIGENT SENSORS SYSTEM

An intelligent sensor is an electronic device for taking measurements of a physical quantity as an electrical signal, it intelligence lie in the ability to check the correct execution of a metrology algorithm, in remote configurability, in its functions relating to the safety, diagnosis, control and communication.

The intelligent sensor can be seen consisting of two parts [10-13]:

1) A measuring chain controlled by microcontroller

2) A bidirectional communication interface with the network, providing the connection of the sensor to a central computer

The communication part reflects all the information collected by an intelligent sensor and allows the user to configure the sensor for operation. It is therefore absolutely essential that this interface be robust and reliable. Figure 1 illustrates the intelligent sensor with its wireless communicating component. A variety of communication interfaces (wireless modules) is available, but not all sensors support these interfaces. The designer must select an interface that provides the best integration of the sensor with the others components of the system taking in our account the costs and the constraints of reliability required for a particular application.

There are others solutions to collect remote measurements such mobile and satellite communications. The main problems related to the quality of communications are: attenuation problems (distance, obstacles, rain ...), interference and multipath. The realization of the systems based on smart sensors dedicated to the applications mentioned in section I, requires the techniques and the protocols that take into account the following constraints [3]:

- The nodes are deployed in high numbers
- At any time, the nodes may be faulty or inhibited
- The topology changes very frequently
- The communication is broadcast

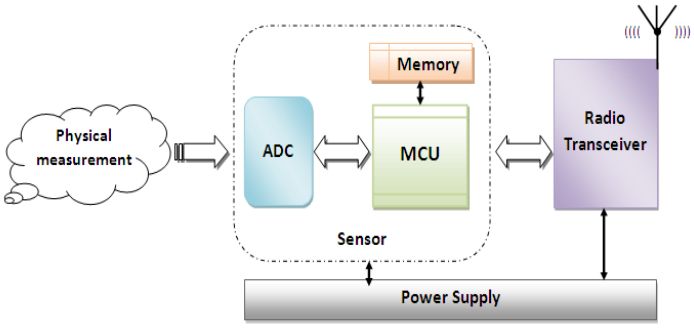


Fig. 1. Block diagram of an intelligent sensor communication

The sensors are limited in energy, in computing capacity, and in memory. In ad-hoc networks, energy consumption was considered as an aim factor but not essential because energy resources can be replaced by the user. These networks are more focused on the QoS than the energy consumption. Contrariwise, in sensor networks, the transmission time and energy consumption are two important performance metrics since generally the sensors are deployed in inaccessible areas.

IV. SENSORS TECHNOLOGY AND OPTIMAL TOPOLOGY

The communication topology of the intelligent sensor systems is divided into two categories:

A. Direct Communication

The intelligent sensors deployed in a capture zone communicate directly with the base station via a radio link as shown in figure 2, the server collect and processes the measurements data and stores it in a database.

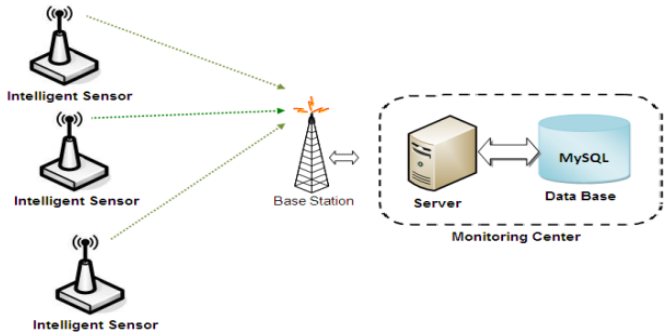


Fig. 2. Direct communication with the monitoring center

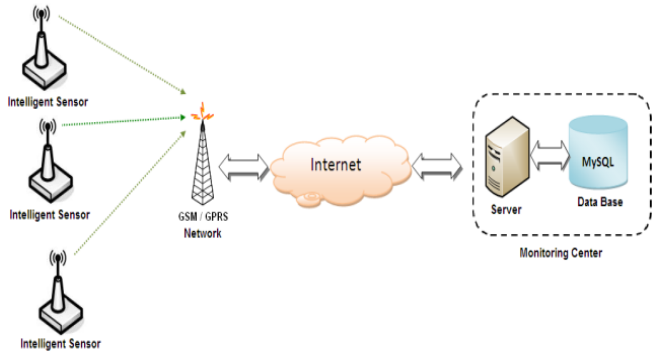


Fig. 3. Indirect communication with the monitoring center

B. Indirect Communication

In this case, the intelligent sensor communicates, via a GPRS network providing Internet connectivity, with the server of the monitoring center as shown in figure 3. With this architecture, it is possible to establish communications for applications that have a wider monitoring area which spreads for kilometers or when the application requires large dimensions.

V. The Comparative Performance Analysis

In this section, we put importance with a comparative study the following wireless protocols: Bluetooth, UWB, ZigBee, ZigBeeIP, Wi-Fi, Wi-Max, GSM/GPRS which correspond to the standards IEEE 802.15.1, 802.15.3, 802.15.4, 802.11a/b/g, 802.16 and 850-900 DCS PCS respectively [14], [15]. Based on the characteristics of each standard, obviously noticed that the UWB, Wi-Fi and Wi-Max protocols provides a higher data rate, while Bluetooth, ZigBee and GPRS provide a lower level.

Contrariwise, Bluetooth, UWB and ZigBee are intended for WPAN communication due to their range of coverage which reaches 10 m, while Wi-Fi is oriented WLAN with a range of about 100 m. However, Wi-Max and GPRS have a coverage radius of a few tens of kilometers for a WMAN network. Table II summarizes the main differences between the mentioned protocols.

TABLE II. Comparison of Characteristics for Different Wireless Protocols

Protocols	**Bluetooth** [2], [14], [17], [18]	**UWB** [14], [19]	**ZigBee/IP** [2], [14], [17-23]	**Wi-Fi** [1], [2], [14], [24], [25]	**Wi-Max** [17], [25-28]	**GSM/GPRS** [29-33]
Frequency band	2.4 GHz	3.1-10.6 GHz	868/915 MHz; 2.4 GHz	2.4; 5 GHz	2.4; 5.1- 66 GHz	850/900; 1800/1900 MHz
Max signal rate	720 Kb/s	110 Mb/s	250 Kb/s	54 Mb/s	35-70 Mb/s	168 Kb/s
Nominal range	10 m	10-102 m	10 - 1000 m	10-100 m	0.3-49 Km	2-35 Km
Nominal TX power	0 - 10 dBm	-41.3 dBm/MHz	-25 - 0 dBm	15 - 20 dBm	23 dBm	0-39 dBm
Number of RF channels	79	(1-15)	1/10; 16	14 (2.4 GHz) 64 (5 GHz)	4;8 10;20	124
Channel bandwidth	1 MHz	0.5- 7.5 GHz	0.3/0.6 MHz; 2 MHz	25-20 MHz	20;10 MHz	200 kHz
Modulation type	GFSK, CPFSK, 8-DPSK, π/4-DQPSK	BPSK, PPM, PAM, OOK, PWM	BPSK QPSK, O-QPSK	BPSK, QPSK, OFDM, M-QAM	QAM16/64, QPSK, BPSK, OFDM	GMSK, 8PSK
Spreading	FHSS	DS-UWB, MB-OFDM	DSSS	MC-DSSS, CCK, OFDM	OFDM, OFDMA	TDMA, DSSS
Basic cell	Piconet	Piconet	Star	BSS	Single-cell	Single-cell
Extension of the basic cell	Scatternet	Peer-to-Peer	Cluster tree, Mesh	ESS	PTMP, PTCM, Mesh	Cellular system
Max number of cell nodes	8	236	> 65000	2007	1600	1000
Encryption	E_0 stream cipher	AES block cipher (CTR, counter mode)	AES block cipher (CTR, counter mode)	RC4 stream cipher (WEP), AES block cipher	AES-CCM cipher	GEA, MS-SGSN, MS-host

Authentication	Shared secret	CBC-MAC (CCM)	CBC-MAC (ext. of CCM)	WPA2 (802.11i)	EAP-SIM, EAP-AKA, EAP-TLS or X.509	PIN; ISP; Mobility Management (GSM A3); RADIUS
Data protection	16-bit CRC	32-bit CRC	16-bit CRC	32-bit CRC	AES based CMAC, MD5-based HMAC, 32-bit CRC	GPRS-A5 Algorithm
Success metrics	Cost, convenience	Throughput, power, cost	Reliability, power, cost	Speed, Flexibility	Throughput, Speed, Range	Range, Cost, Convenience,
Application focus	Cable replacement	Monitoring, Data network,	Monitoring, control	Data network, Internet, Monitoring,	Internet, Monitoring, Network Service,	Internet, Monitoring, control

VI. Characteristics of Wireless Communication Protocols

We present in this section the different metrics to measure the performance of a wireless protocol.

A. Network Size

The size of the GPRS network can be balanced according to the interference level, the size of data packets during traffic, the transmission protocols implemented and the number of users connected to the GSM voice services, this influences the number of GPRS open sessions which can reach 1000 to a single cell. ZigBee star network take the first rank for the maximum number of nodes that exceeds 65000, in second place there is the Wi-Fi network with a number 2007 of nodes in the BSS structure, while the Wi-Max network has a size of 1600 nodes, UWB allows connection for 236 nodes in the piconet structure, finally we found the Bluetooth which built its piconet network with 8 nodes. All these protocols have a provision for more complex network structures built from basic cells which can be used to extend the size of the network.

B. Transmission Time

The transmission time depends on the data rate, the message size, and the distance between two nodes. The formula of transmission time in (µs) can be described as follows:

$$T_{tx} = \left(N_{data} + \left(\frac{N_{data}}{N_{maxPld}} \times N_{ovhd} \right) \right) \times T_{bit} + T_{prop} \qquad (1)$$

N_{data} the data size

N_{maxPld} the maximum payload size

N_{ovhd} the overhead size

T_{bit} the bit time

T_{prop} the propagation time between two nodes to be neglected in this paper

The typical parameters of the different wireless protocols used to evaluate the time of transmission are given in Table III.

TABLE III. Typical Parameters Of Wireless Protocols

Protocol	Max data rate (Mbit/s)	Bit time (µs)	Max data payload (bytes)	Max overhead (bytes)	Coding efficiency[+] (%)
Bluetooth	0.72	1.39	339 (DH5)	158/8	94.41
UWB	110	0.009	2044	42	97.94
ZigBee	0.25	4	102	31	76.52
Wi-Fi	54	0.0185	2312	58	97.18
Wi-Max	70	0.0143	2700	40	98.54
GPRS	0.168	5.95	1500[*]	52[*]	80.86

[+] Where the data is 10 Kbytes. [*] For TCP/IP Protocol

From the figure 4, it is noted that the transmission time for the GSM/GPRS is longer than the others, due to its low data rate (168 Kb/s) and its long range reasons, while UWB requires less transmission time compared to the others because its important data rate.

It clearly shows that the required transmission time is proportional to the data payload size N_{data} and it is not proportional to the maximum data rate.

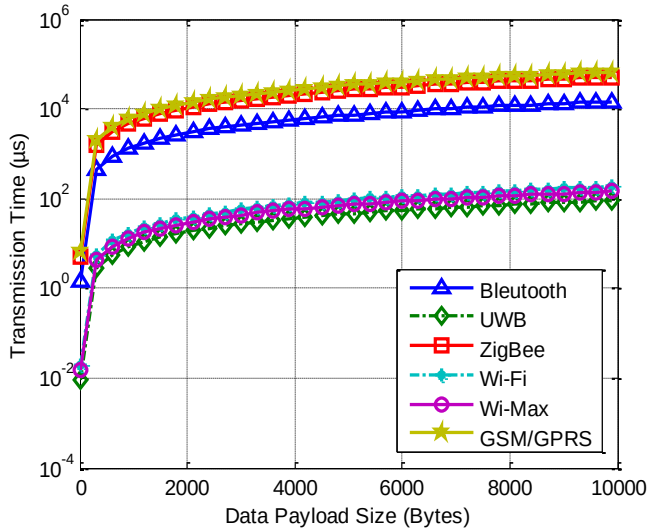


Fig. 4. Comparison of transmission time relative to the data size

C. Transmission power and range

In wireless transmissions, the relationship between the received power and the transmitted power is given by the Friis equation as follows [1], [33], [36-40]:

$$\frac{P_r}{P_t} = G_t G_r \left(\frac{\lambda}{4\pi D}\right)^2 \qquad (2)$$

P_t the transmitted power
P_r the received power
G_t the transmitting omni basic antenna gain
G_r the receiving antenna gain
D the distance between the two antennas
λ the wavelength of the signal

From equation (2) yields the formula the range of coverage as follows:

$$D = \frac{1}{\frac{4\pi}{\lambda}\sqrt{\frac{P_r}{P_t G_t G_r}}} \qquad (3)$$

We note that as the frequency increases, the range decreases. The figure 5 shows the variation of signal range based on the transmission frequency for a fixed power. The most revealing characteristic of this graph is the non-linearity. The signals of GSM/GPRS with 900MHz propagate much better than ZigBee, Wi-Fi, Bluetooth with 2.4GHz and UWB with 3.1GHz vice to vice coverage area.

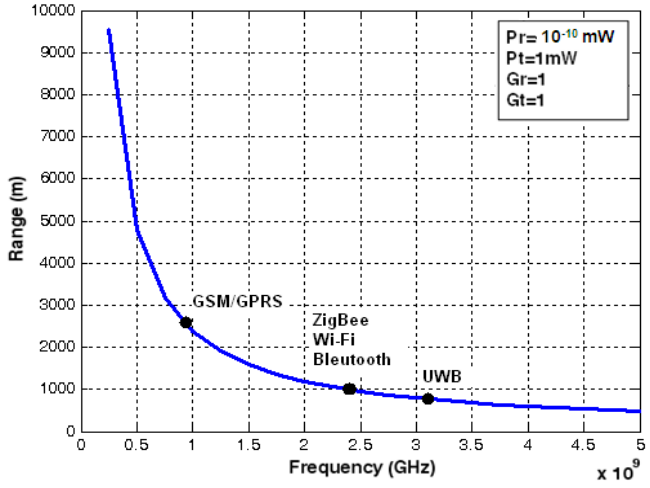


Fig. 5. Range evolution according to the transmission frequency

D. Energy consumption

The energy consumption for intelligent sensor involves three steps: acquisition, communication, computation and data aggregation. This consumption in the acquisition operation depends on the nature of the application [3]. Data traffic, particularly in the transmission, consumes more energy than the other operations. It also depends on the distance between the transmitter and receiver [4], [5].

The model governing the energy consumption E(p) of an intelligent sensor p depending on the communication range d(p) is given as follows:

$$E(p) = k.d^{\alpha}(p) + E_d \quad \alpha \geq 2 \qquad (4)$$

k the packet size
α the signal attenuation coefficient
E_d the transmission energy costs

According to the radio energy model, [6], [38-44] the transmission power of a k bit message to a distance d is given by:

$$E_{TX}(k,d) = \begin{cases} k.\varepsilon_{fs}.d^2 + k.E_{Elec} & d < d_0 \\ k.\varepsilon_{amp}.d^4 + k.E_{Elec} & d \geq d_0 \end{cases} \qquad (5)$$

$$d_0 = \sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{amp}}} \qquad (6)$$

E_{Elec} electronic energy
ε_{fs}, ε_{amp} amplification energy

The electronic energy E_{Elec} depends on several factors such as digital coding, modulation, filtering, and signal propagation, while the amplifier energy depends on the distance to the receiver and the acceptable bit error rate. If the message size and the range of communication are fixed, then if the value of α grow, the required energy to cover a given distance increase also.

The figure 6 illustrates the evolution of the energy consumption for ZigBee protocol based on the signal range. We can say that an increase in data packet size allows then an increase of the transmission energy. The equations (4) and (5) can be generalized for the all wireless mentioned protocols. The simulation parameters are given in table IV.

TABLE IV. THE SIMULATION PARAMETERS

Parameters	Value
E_{Elec}	50 nJ/bit
ε_{fs}	10 pJ/bit/m^2
ε_{amp}	0.0013 pJ/bit/m^4

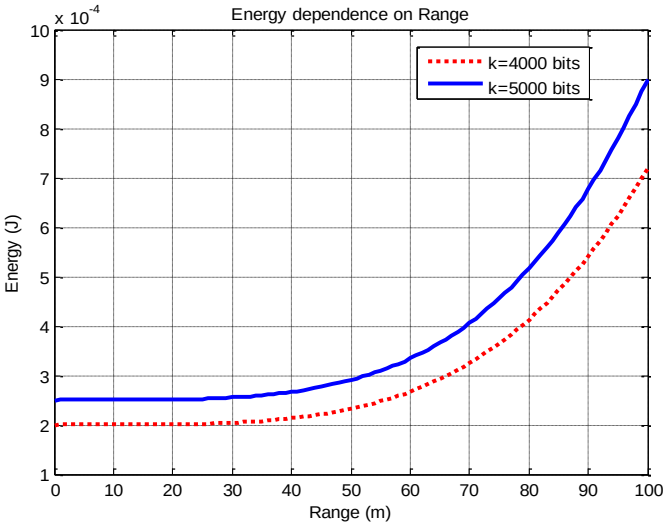


Fig. 6. The energy consumption depending on the signal range

The predicted received power by an intelligent sensor for each data packet according to the communication range d is given by the Two-Ray Ground and the Friss free space models [3], [35], [40] as follows:

$$P_r(d) = \begin{cases} \dfrac{P_t G_t G_r \lambda^2}{(4\pi d)^2 L} & d < d_c \\ \dfrac{P_t G_t G_r h_t^2 h_r^2}{d^4} & d \geq d_c \end{cases} \quad (7)$$

$$d_c = \frac{4\pi\sqrt{L} h_r h_t}{\lambda} \quad (8)$$

L the path loss
h_t the height of the transmitter antenna
h_r the height of the receiver antenna
d the distance between transmitter and receiver

The figure 7 shows the evolution of the reception power based on the signal range for the different studied protocols for a fixed data packet size:

TABLE V. THE SIMULATION PARAMETERS

Parameters	Value
L	1
G_t=G_r	1
h_t=h_r	1.5 m

Protocols	Transmitted Power (Watt)
Bluetooth	0.1
UWB	0.04
ZigBee	0.0063
Wi-Fi	1
Wi-Max	0.25
GSM/GPRS	2

According to this figure, it is noted that when the distance between the transmitter and the receiver increases, the received power decreases, this is justified by the power loss in the path. The ZigBee, UWB and Bluetooth have low power consumption while Wi-Max, Wi-Fi and GPRS absorb more power due to theirs high communication range reason.

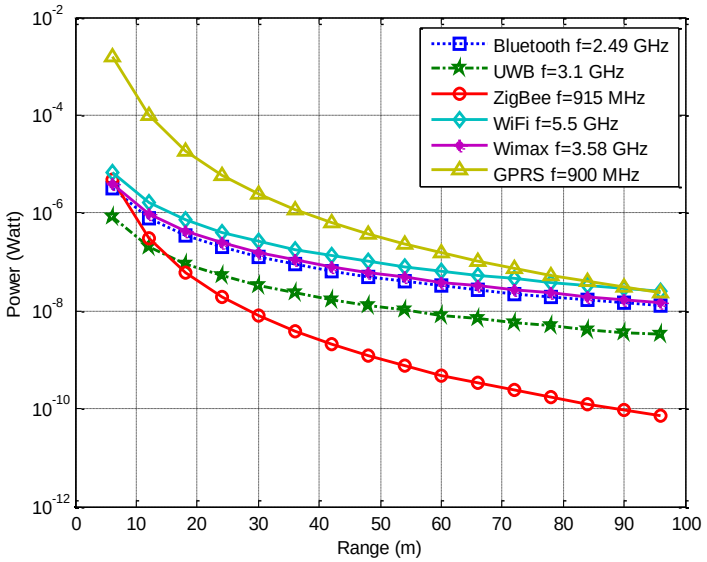


Fig. 7. The received power depending on the signal range with fixed message size

E. Chipset power consumption

To compare practically the power consumption, we are presents in the table VI the detailed representative characteristics of particular chipset for each protocol [44-49]. The figure 8 shows the consumption power in (mW) for each protocol. Obviously we note that Bluetooth and ZigBee consume less power compared to UWB, Wi-Fi, Wi-Max and a GPRS connection. The difference between the transmission power and reception power for the protocols GPRS and Wi-Max is justified by the power loss due to the attenuation of the

signal in the communication path since both of these protocols have a large coverage area.

TABLE VI. POWER CONSUMPTION CHARACTERISTICS OF CHIPSETS

Protocols	Chipset	V_{DD} (volt)	I_{TX} (mA)	I_{RX} (mA)	Bit rate (Mb/s)
Bluetooth	BlueCore2	1.8	57	47	0.72
UWB	XS110	3.3	~227	~227	114
ZigBee	CC2430	3.0	24.7	27	0.25
Wi-Fi	CX53111	3.3	219	215	54
Wi-Max	AT86 RF535A	3.3	320	200	70
GSM/GPRS	SIM300	3	350*	230*	0.164*

* For GSM 900 DATA mode, GPRS (1 Rx,1 Tx)

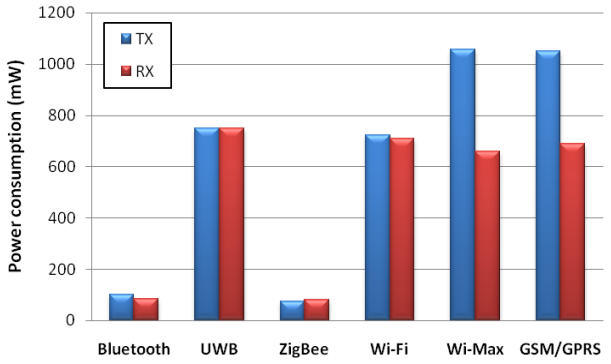


Fig. 8. Comparison of chipset power consumption for each protocol

Based on the data rate of each protocol, the normalized energy consumption in (mJ/Mb) is shown in the figure 9, shows clearly in this figure that the UWB, Wi-Fi and Wi-Max have better energy efficiency. In summary, we can say that Bluetooth and ZigBee are suitable for low data rate applications with a limited battery power, because of their low energy consumption which promotes a long lifetime. Contrariwise for implementations of high data rate, UWB, Wi-Fi and Wi-Max would be the best solution due to their low normalized energy consumption. While for monitoring and surveillance applications with low data rate requiring large area coverage, GPRS would be an adequate solution.

F. Bit error rate

The transmitted signal is corrupted by white noise AWGN (Additive White Gaussian Noise) to measure the performance of the digital transmissions (OQ-B-Q-PSK, 4PAM, 16QAM, GMSK, GFSK, 8DPSK, 8PSK and OFDM), seen in the table II, by calculating the bit error probability. The purpose of a modulation technique is not only the transfer of a data packet by a radio channel, but also achieves this operation with a better quality, energy efficiency and less bandwidth as possible.

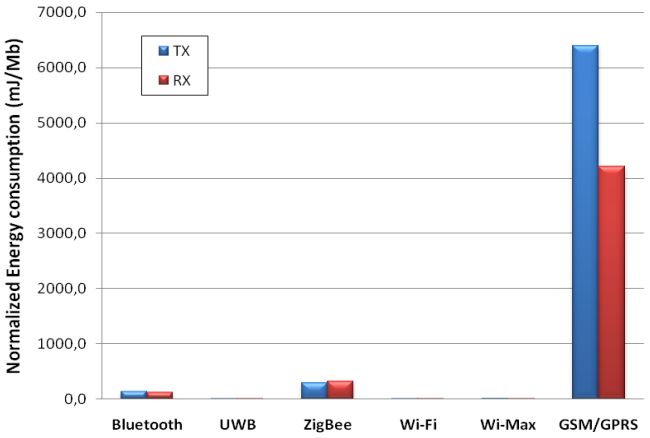


Fig. 9. Comparing the chipset normalized energy consumption for each protocol

The bit error rate is a very good way to measure the performance of the modulation used by a communication system and therefore helps to improve its robustness. It is calculated by the following formula:

$$BER = \frac{N_{Err}}{N_{TXBits}} \quad (9)$$

N_{Err} the number of errors
N_{TXBits} the number of transmitted bits

The figure 10 shows the BER of the differents modulations used in wireless technologies mentioned above according on signal to noise ratio E_b/N_0.

The BER for all systems decreases monotonically with increasing values of E_b/N_0, the curves defining a shape similar to the shape of a waterfall [36], [38]. The BER for QPSK and OQPSK is the same as for BPSK. We note that the higher order modulations exhibit higher error rates which thus leads to a compromise with the spectral efficiency.

QPSK and GMSK seem the best compromise between spectral efficiency and BER followed by other modulations. These two robust modulations are used in Wi-MAX, ZigBee, Wi-Fi and in GPRS network, can be employed in the noisy channels and in the noisy environments. However, because of their sensitivity to noise and non-linearities, the modulations 4PAM and 8DPSK remain little used compared to other modulations.

Concerning the QAM modulation, it uses more efficiently the transmitted energy when the number of bits per symbol increases; this provides a better spectral efficiency and a high bit rate. As for the frequency hopping FSK modulations, the increase of the symbols will enable reduction of the BER but also increase the spectral occupancy. The main fault of these FSK modulations is their low spectral efficiency.

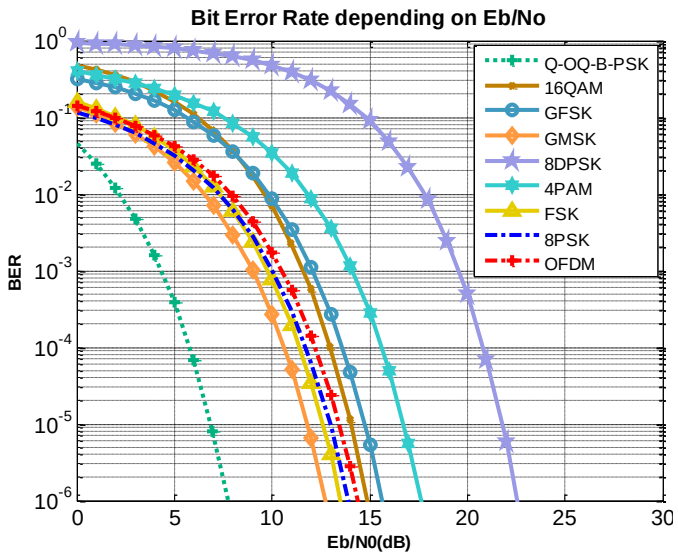


Fig. 10. Bit Error Rate for differents modulations

On the other side, the GMSK modulation has been developed in order to increase the spectral efficiency [50]. It has a satisfactory performance in terms of BER and noise resistance. This modulation is applied in the data transmission systems (MODEM), in The GSM networks [9], [35], [37], [39], [41]. The table VII gives the values of E_b/N_0 which cancel the BER for each modulation. Furthermore, the lower bit error probability is obtained to the detriment of the number of users. We must investigate the relationship between the transmission quality and the number of users served [50].

TABLE VII. E_b/N_0 VALUES WHICH CANCELS BER FOR THE DIFFERENT MODULATIONS

Modulation	E_b/N_0 (dB)	B.E.R
B-OQ-QPSK	7,8	10^{-6}
GMSK	12,7	10^{-6}
FSK	13,3	10^{-6}
8PSK	13,8	10^{-6}
OFDM	14,3	10^{-6}
16QAM	14,8	10^{-6}
GFSK	15,7	10^{-6}
4PAM	17,6	10^{-6}
8DPSK	22,6	10^{-6}

G. Data coding efficiency

The coding efficiency can be calculated from the following formula:

$$P_{cdeff} = 100 \times \frac{N_{data}}{\left(N_{data} + \left(\frac{N_{data}}{N_{maxPld}} \right) \times N_{ovhd} \right)} \qquad (10)$$

Based on the figure 11, the coding efficiency increases when the data size increase. For small data size, Bluetooth and ZigBee is the best solution while for high data sizes GPRS, UWB, Wi-Max and Wi-Fi protocols have efficiency around 94%.

In the applications point of view, for the automation industrial systems based on intelligent sensors, since most data monitoring and industrial control have generally a small size, such the pressure or the temperature measurements that don't pass 4 bytes and that don't require an important data rate, Bluetooth, ZigBee and GPRS can be a good choice due to their coding efficiency and their low data rate. On the other hand, for applications requiring a large cover zone as the borders monitoring, the persons tracking or the environmental monitoring or the event detection, GPRS and Wi-Max are an adequate solution, whereas for the multimedia applications requiring an important data rate such the video monitoring, Wi-Fi, UWB and Wi-Max form a better solution.

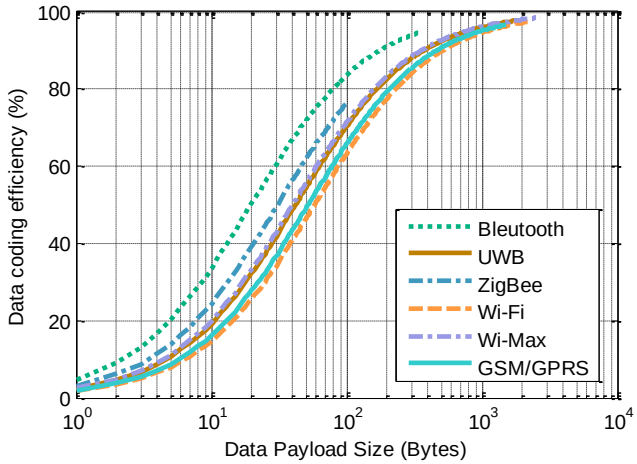


Fig. 11. Coding efficiency depending on data size

VII. CONCLUSION

We have presented in this paper a comparative performance analysis of six wireless protocols: Bluetooth, UWB, ZigBee, Wi-Fi, Wi-Max and GSM/GPRS. However, it exists others wireless protocols as 6LoWPAN, DASH, HiperLAN…We have chosen therefore to land just the most popular ones. A quantitative evaluation of the transmission time, the data coding efficiency, the bite error rate, and the power and the energy consumption in addition of the network size permitted us to choose the best protocol which is suitable for an application based on intelligent sensor.

Furthermore, the adequacy of these protocols is influenced strongly by many others factors as the network reliability, the link capacity between several networks having different protocols, the security, the chipset price, the conformity with the application and the cost of installation that must be taking in consideration. Facing the fact that several types of wireless technologies can coexist in a capture environment, the challenge which requires is to develop a gateway (multi-standard transceiver) that enables the data exchange between these heterogeneous infrastructures with a good quality of service. This approach would allow the implementation of solutions for maintaining and for monitoring while minimizing the necessary resources and avoiding the costs associated to the compatibility testing. Solving this challenge is a

perspective and a continuation of this work. It turns out that the choice of a modulation type is always determined by the constraints and the requirements of the application. The BER is a parameter which gives an excellent performance indication of a radio data link.

REFERENCES

[1] Jean-Paul M., G. Linmartz's, "Wireless Communication, The Interactive Multimedia CD-ROM", Baltzer Science Publishers, P.O.Box 37208, 1030 AE Amsterdam, ISSN 1383 4231, Vol. 1 (1996), No.1

[2] Helen Fornazier et al., "Wireless Communication : Wi-Fi, Bluetooth, IEEE 802.15.4, DASH7", ROSE 2012 ELECINF344 / ELECINF381, Télécom ParisTech, web site : http://rose.eu.org/2012/category/admin

[3] Lehsaini Mohamed, "Diffusion et couverture basées sur le clustering dans les réseaux de capteurs : application à la domotique ", Thèse de Doctorat, Université de Franche-Comté Besançon, U.F.R Sciences et Techniques, École Doctorale SPIM, Juillet 2009

[4] Trevor Pering et al., CoolSpots: "Reducing the Power Consumption of Wireless Mobile Devices with Multiple Radio Interfaces", ACM 1-59593-195-3/06/0006, MobiSys'06, June 19–22, 2006, Uppsala, Sweden

[5] Travis Collins et al., "A Green Approach to a Multi-Protocol Wireless Communications Network", Major Qualifying Project to obtain the Degree in Bachelor of Science in Electrical and Computer Engineering, Faculty of Worcester Polytechnic Institute, University of Limerick 2011, http://www.wpi.edu/Academics/Projects.

[6] Wendi B. Heinzelman et al., "An Application-Specific Protocol Architecture for Wireless Microsensor Networks", IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, VOL. 1, NO. 4, OCTOBER 2002

[7] Baghouri Mostafa, Chakkor Saad, Hajraoui Abderrahmane, "Fuzzy logic approach to improving Stable Election Protocol for clustered heterogeneous wireless sensor networks", Journal of Theoretical and Applied Information Technology, Vol. 53 No.3, July 2013

[8] El Ahmadi Cheikh, Chakkor Saad et al, "New Approach to Improving Lifetime in Heterogeneous Wireless Sensor Networks Based on Clustering Energy Efficiency Algorithm", Journal of Theoretical and Applied Information Technology, Vol. 61 No.2, March 2014

[9] Crepin Nsiala Nzeza, " RÉCEPTEUR ADAPTATIF MULTI-STANDARDS POUR LES SIGNAUX A ÉTALEMENT DE SPECTRE EN CONTEXTE NON COOPÉRATIF ", thèse de Doctorat, UNIVERSITÉ de Bretagne Occidentale, Juillet 2006

[10] Guillaume Terrasson, "CONTRIBUTION A LA CONCEPTION D'EMETTEUR- RECEPTEUR POUR MICROCAPTEURS AUTONOMES", thèse de Doctorat, UNIVERSITÉ BORDEAUX 1, Novembre 2008

[11] Creed Huddleston, "Intelligent Sensor Design Using the Microchip dsPIC (Embedded Technology)", Newnes 2007

[12] Elmostafa Ziani, "CONCEPTION ET REALISATION D'UN INSTRUMENT ULTRASONORE INTELLIGENT DEDIE A LA MESURE DE DEBITS D'ECOULEMENT A SURFACE LIBRE", Thèse de doctorat en cotutelle, Université abdelmalek essaadi, Faculté des sciences et techniques de Tanger, Université de paris 13 Villetaneuse, Institut universitaire de technologie de Saint- denis 2005

[13] Fei Hu, Qi Hao, "Intelligent Sensor Networks: The Integration of Sensor Networks", Signal Processing and Machine Learning, CRC Press 2012

[14] Jin-Shyan Lee et al., "A Comparative Study of Wireless Protocols: Bluetooth, UWB, ZigBee", and Wi-Fi, The 33rd Annual Conference of the IEEE Industrial Electronics Society (IECON), Taipei, Taiwan, November 5-8, 2007

[15] Adil Koukab et al., "A GSM-GPRS/UMTS FDD-TDD/WLAN 802.11a-b-g Multi-Standard Carrier Generation System", IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 41, NO. 7, JULY 2006

[16] Jin-Shyan Lee, "Performance Evaluation of IEEE 802.15.4 for Low-Rate Wireless Personal Area Networks", IEEE Transactions on Consumer Electronics, Vol. 52, No. 3, AUGUST 2006

[17] Klaus Gravogl et al., "Choosing the best wireless protocol for typical applications", 2nd Workshop on Ultra-low Power Wireless Sensor Networks (WUPS 2011) February 2011, Como, Italy, http://geodes.ict.tuwien.ac.at/PowerSavingHandbook/

[18] Z. Mammeri, " Réseaux sans fils Caractéristiques et principaux standards ", M1 Info Cours de Réseaux, IRIT, Université Paul Sabatier, Toulouse http://www.irit.fr/~Zoubir.Mammeri/Chap6WLAN.pdf

[19] Ghobad Heidari, "WiMedia UWB: Technology of Choice for Wireless USB and Bluetooth", edition John Wiley & Sons Ltd 2008, ISBN 978-11-470-51814-2 (HB)

[20] Ms. Dharmistha, D. Vishwakarma, "IEEE 802.15.4 and ZigBee: A Conceptual Study", International Journal of Advanced Research in Computer and Communication Engineering, ISSN : 2278 – 1021, Vol. 1, Issue 7, September 2012

[21] Vaddina Prakash Rao, "The simulative Investigation of Zigbee/IEEE 802.15.4", Master Thesis of Science, DRESDEN UNIVERSITY OF TECHNOLOGY, FACULTY OF ELECTRICAL ENGINEERING AND INFORMATION TECHNOLOGY, Department of Electrical Engineering and Information Technology, Chair of Telecommunications, September, 2005

[22] http://www.zigbee.org/Specifications/ZigBeeIP/Overview.aspx

[23] Reen-Cheng Wang et al., "Internetworking Between ZigBee/802.15.4 and IPv6/802.3 Network", ACM 978-1-59593-790-2/07/0008, IPv6'07, August 31, 2007, Kyoto, Japan.

[24] Aurélien Géron, "WIFI PROFESSIONNEL La norme 802.11, le déploiement, la sécurité ", 3ème édition DUNOD

[25] Bhavneet Sidhu et al., "Emerging Wireless Standards - WiFi, ZigBee and WiMAX", World Academy of Science, Engineering and Technology 25 2007

[26] Michèle Germain, "WiMAX à l'usage des communications haut débit", Forum atena, lulu.com, Paris, 2009

[27] Loutfi Nuaymi, "WiMAX : Technology for Broadband Wireless Access", Wiley, 2007

[28] Marwa Ibrahim et al., "Performance investigation of Wi-Max 802.16m in mobile high altitude platforms", Journal of Theoretical and Applied Information Technology, 10th June 2013. Vol. 52 No.1

[29] Xavier Lagrange et al., " Réseaux GSM : des principes à la norme ", Éditions Hermès Sciences, 2000, ISBN 2-7462-0153-4

[30] Timo Halonen et al., "GSM, GPRS Performance and EDGE, Evolution Towards 3G/UMTS", Second Edition, John Wiley & Sons Ltd 2003, ISBN 0-470-86694-2

[31] Brahim Ghribi, Luigi Logrippo, "Understanding GPRS: the GSM packet radio service", Elsevier Computer Networks 34 (2000) 763 - 779

[32] Christian Bettstetter et al., "GSM phase 2+general packet radio service GPRS: architecture, protocols, and air interface", IEEE Communications Surveys, Third Quarter 1999, vol. 2 no. 3, http://www.comsoc.org/pubs/surveys

[33] Joseph Ho et al., "Throughput and Buffer Analysis for GSM General Packet Radio Service (GPRS)", Wireless Communications and Networking Conference New Orleans, LA, 1999. WCNC. 1999 IEEE, Pages 1427 - 1431 vol.3, ISSN: 1525-3511, ISBN: 0-7803-5668-3

[34] Constantine A. Balanis, "Antenna Theory: Analysis and Design" (2nd edition), John Wiley and Sons, Inc 1997

[35] François de Dieuleveult, Olivier Romain, " ÉLECTRONIQUE APPLIQUÉE AUX HAUTES FRÉQUENCES " Principes et applications, 2e édition, Dunod, Paris 2008, ISBN 978-2-10-053748-8

[36] DR. Kamilo Feher, "Wireless Digital Communications (Modulation & Spread spectrum Applications)", PHI Learning, Prentice Hall PTR, 1995

[37] Simon Haykin, "Communication Systems", 4Th Edition with Solutions Manual, John Wiley and Sons, Inc 2001.

[38] Proakis, J. G., "Digital Communications", 3rd edition, New York, McGraw-Hill, 1995.

[39] Sklar, B., "Digital Communications: Fundamentals and Applications", Englewood Cliffs, NJ, Prentice-Hall, 1988

[40] Rappaport T.S., "Wireless Communications Principles and Practice", 2nd Edition, Prentice Hall, 2001.

[41] Andreas F. Molisch, "WIRELESS COMMUNICATIONS", John Wiley & Sons Ltd., Second Edition 2011, ISBN: 978-0-470-74187-0

[42] Prakash C. Gupta, "Data Communications and Computer Networks",

PHI Learning, Prentice-Hall of India 2006, ISBN 81-203-2846-9

[43] Nitin Mittal et al., "IMPROVED LEACH COMMUNICATION PROTOCOL FOR WSN", NCCI 19-20 March 2010, National Conference on Computational Instrumentation CSIO Chandigarh, INDIA

[44] Cambridge Silicon Radio, BlueCore2-External Product Data Sheet. Cambridge, UK, Aug. 2006.

[45] [45] Freescale, XS110 UWB Solution for Media-Rich Wireless Applications. San Diego, CA, Dec. 2004.

[46] Chipcon, CC2430 Preliminary Data Sheet (rev. 1.03). Oslo, Norway, 2006.

[47] Conexant, Single-Chip WLAN Radio CX53111. Newport Beach, CA, 2006.

[48] SIMCOM Ltd, SIM300 Hardware Specification, 27th Dec 2005.

[49] ATMEL, WiMax Transceiver 802.16-2004, AT86RF535A Preliminary Data Sheet, 2006

[50] David K. Asano, Subbarayan Pasupathy, "Optimization of Coded GMSK Systems", IEEE Transactions on Information Theory, VOL. 48, NO. 10, OCTOBER 2002

[51] "Guide of MATLAB" 7.8.0 (R2009a), www.mathworks.com

Multi-Domain Modeling and Simulation of an Aircraft System for Advanced Vehicle-Level Reasoning Research and Development

F. Khan, O. F. Eker, T. Sreenuch
Integrated Vehicle Health Management
Centre Cranfield University
Bedford MK43 0AL, UK

A. Tsourdos
Engineering Science Division
Cranfield University
Bedford MK43 0AL, UK

***Abstract*—In this paper, we describe a simulation based health monitoring system test-bed for aircraft systems. The purpose of the test-bed is to provide a technology neutral basis for implementing and evaluation of reasoning systems on vehicle level and software architecture in support of the safety and maintenance process. This simulation test-bed will provide the sub-system level results and data which can be fed to the VLRS to generate vehicle level reasoning to achieve broader level diagnoses. This paper describes real-time system architecture and concept of operations for the aircraft major sub-systems. The four main components in the real-time test-bed are the aircraft sub-systems (e.g. battery, fuel, engine, generator, heating and lighting system) simulation model, fault insertion unit, health monitoring data processing and user interface. In this paper, we adopted a component based modelling paradigm for the implementation of the virtual aircraft systems. All of the fault injections are currently implemented via software. The fault insertion unit allows for the repeatable injection of faults into the system. The simulation test-bed has been tested with many different faults which were undetected on system level to process and detect on the vehicle level reasoning. This article also shows how one system fault can affect the overall health of the vehicle.**

Keywords*—Intelligent Reasoning; Finite State Machines; Aircraft System Simulation; Multi-Physics; Real-Time Simulation, VLRS (Vehicle Level Reasoning System); HM (Health Management)

I. Introduction

A Vehicle Level Reasoning System (VLRS) aids in enhancing the safety of the aircraft. Such systems comprise of various units (sub-system reasoner) that monitor related components for functional status and relay back operational status to the entities of interest. Thus, a primary function of the VLRS, see Figure 1, is to deduce the overall operational health of the aircraft.

The VLRS takes data/results input from several sub-systems and processes this information to provide overall vehicle health status[1],[2],[3]. However the major challenge is the sub-system level data and results are not available on vehicle level (includes several sub-systems with connected physics).

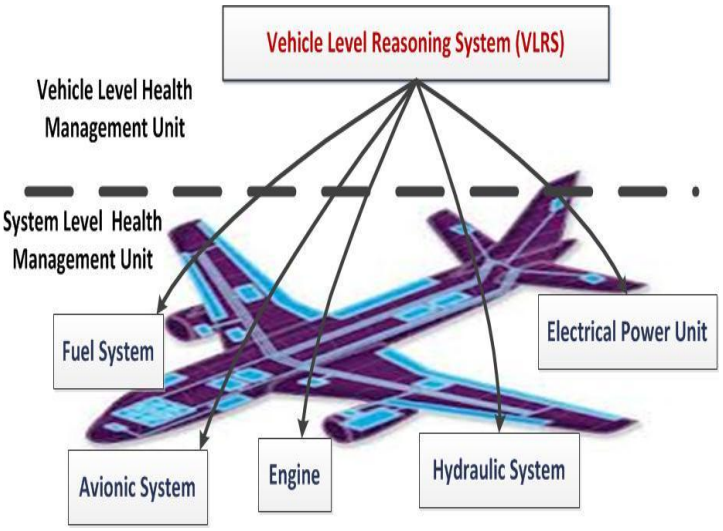


Fig. 1. VLRS overview.

One of the objectives of this simulation test-bed is to demonstrate VLRS, Artificial Intelligence Exchange and Service Tie to All Test Environments (AI-ESTATE)[4]and Open System Architecture for Condition Based Maintenance (OSA-CBM)[5]. The aim of the current work is to develop a simulation test-bed that emulates hardware for a practical aerospace related application and implement a health monitoring system for the test-bed [6].

With the lack of large scale diagnostic test-beds and in order to meet the complexity of the aerospace applications, we have developed the diagnostic test-bed with the following goals in mind:

- Provide a data and results sub-system to create a VLRS
- Provide the capability to perform testing of diagnostic algorithms by manually or algorithmically inserting faults, and
- Provide a technology neutral basis for implementing and evaluation diagnostic systems and software architecture, tosupport the condition based maintenance (CBM) process.

In this case, a system representation of an aircraft's basic core sub-system (i.e. batter, fuel, engine, generate, heating and lighting system) is chosen as an example for the simulation development. These sub-systems and their associated health monitoring algorithm will then be used to develop a VLRS (Vehicle Level Reasoning System) for demonstration.

This simulation is considered to be a good candidate for a test-bed to be used for data processing architecture evaluation purposes, i.e. being composed of many systems/components to be monitored and many sensors that generate data. These data can be collected and used to generate diagnostics results at vehicle level across the aircraft systems.

The system that has been used in this simulation can provide multi-physics data such as electric, temperature, fuel flow and pressure which will enable us to insert faults in the fuel system and different electrical systems.

Generally aircraft's all major systems are very complex in nature, Hence the modelled system in this simulation are very complex in their design,therefore this simulation has been created by the basic design of these systems by aiming to create a system level data to perform a testing of Vehicle level reasoners simulation platform.

II. Litreature Review

A. Detection and Diagnostic System

Detection and diagnosis can be achieved manually, by rule-based systems, mathematical or other learning or model based techniques. Fault detection and diagnosis in systems have been widely used in commercial industry over the past few decades[7], [8]. The diagnostics algorithms can be based on different types of measurement depending on the systems and applications, for example, the electrical currents in Motor Current Signal Analysis (MCSA) [9] and accelerations in Vibration Analysis [10]. The purpose of these methods is to detect and diagnose faults at an early stage and therefore allow contingency plans to be put into place before the problems worsen.

Historically, troubleshooting has been a major element of the maintenance strategy for mechanical equipment of any kind. The traditional diagnostics monitoring equipment detects any abnormal behaviour and triggers a ground based test or troubleshooting activity. Nearly all systems, especially more complex aerospace systems, fall short of the ideal system that could accurately and unambiguously drive replacement or repair actions with no additional testing required. Inherent diagnostic ambiguity and conditions that lead to false alarms results in extensive troubleshooting, parts swapping and shotgun maintenance which increases in turnaround time and maintenance costs[11]. This of course has an impact on further development for the diagnostic reasoners, initiating several different techniques such as Model Based Reasoning (MBR) or data driven methods[12].

NASA was an early contributor to vehicle level reasoning systems. In 2004, NASA uploaded Livingstone Version 2 (LV2) software to the EO-1 satellite to test its ability to find and analyse errors in the spacecraft's system,[13].

Sponsors of this project are IVHM Cranfield University, Boeing, BAE System and other IVHM partners

TABLE I. Demonstrates Diagnostics Reasoners and Their Companies

Intelligent Reasoner	Type	Known Applications	Company Information
CMC	Fault propagation modelling	Boeing 777; Primus Epic (business jets, Helicopters)	Honeywell International
TEAMS Toolset	Multi-signal dependency modelling (advanced form modelling)	Consult Company	Qualtech Systems Inc.
eXpress Design Toolset	Dependency modelling (similar to fault propagation modelling)	Consult Company	DSI international
Livingstone	Artificial intelligence based reasoner (mixture of functional and parametric modelling)	DEEP Space One Spacecraft ; Earth observing one (EO-1) satellite	NASA Ames Research Centre
BEAM	Artificial intelligence based reasoner (mixture of functional and parametric modelling)	NASA Deep Space Missions (Voyager, Galileo, Megellan, Cassini and Extreme Ultraviolet explorer	NASA Jet Propulsion Laboratory

Tests, normally performed on the ground, were conducted in flight to automatically detect and diagnose simulated failures in the satellite's instruments and systems. Livingstone provides the opportunity to recover from errors to protect these assets, and continue to achieve mission goals. On this mission, LV2 also monitored another software application that controlled EO-1 to autonomously run its imaging system. If EO-1 did not respond properly to the software control, LV2 detects the error, makes a diagnosis, and sends its analysis to mission control. LV2 compares a model of how the spacecraft's systems and software should perform to the actual performance. If the spacecraft's behaviour differs from the model, then the LV2 reasoners search for the root cause and provide mission controllers suggestions of what may have gone wrong. Actually very few aircraft have VLRS built in, even those VLRS are based on a basic detection and pattern recognition diagnostics system. These VLRS have been designed concurrently with the aircraft and do not incorporate plug and play facilities. Therefore, including any further sub-systems in the VLRS is not feasible. The literature review shows very little information about VLRS implementation in military and no implementation in civil aircraft.

III. Challenges To Vlrs Modelling

VLRS is there to detect and predict faults and failures at the aircraft level. It does this by receiving health information from individual sub-systems and fusing them to derive an overall health status for the aircraft. Generally, the reasoning system is

an artificial intelligence based software application, hardware device or combination of hardware and software whose computational function is to generate conclusions from available knowledge using logical techniques of deduction, diagnosing and prediction or other forms of reasoning [14].

In an aircraft the sub-systems are developed by many different vendors, see Figure 2. Each vendor has their own development and design philosophy and will use the best diagnostic algorithm for their equipment. Such algorithms will produce results that are interfaced to the aircraft system via a communication bus such as ARINC 429. To enable the communication on an ARINC 429 bus the component has to follow the interface standards.

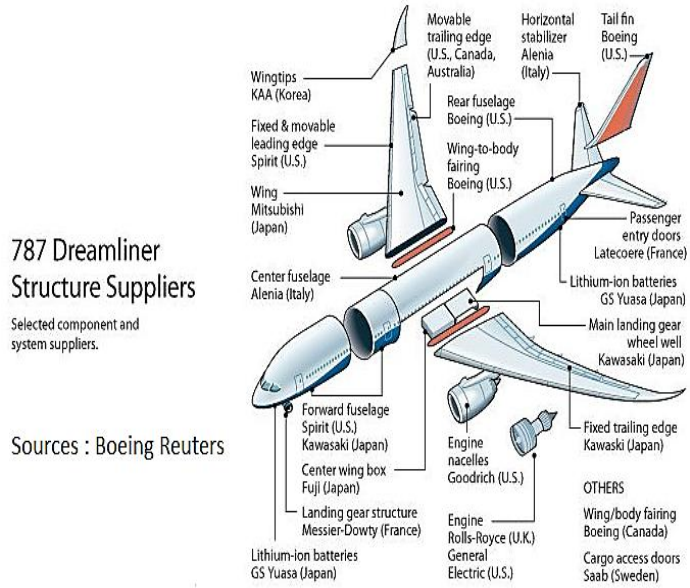


Fig. 2. Boeing 787 an example figure for aircraft and their vendors.

Figure 2 demonstrates the suppliers and parts provided by different vendors for just one aircraft model; in this case it shows the parts made by the different countries and vendors. This shows that in there are major issues that must be dealt with when developing/manufacturing the aircraft. Each component, sub-system and part has been designed by a much defined outlined interface in order to communicate with CMC (Central Maintenance Computer) or sub-system [15]. If this outlines changes it would require component, sub-system vendors and suppliers to change their design as well as to communicate with the rest of the system. It would be unreasonable to require vendors and suppliers to use particular algorithm techniques. Their systems have to go through intensive testing and certification before they can be used in a commercial aircraft; further testing would mean more cost. Consequently, interoperability between the components or sub-systems supplied by different vendors has essentially become one of the major challenges for VLRS. With each component being specific in its nature, there exists a need for a common communication vocabulary that allows for health status communication between the components and the VLRS.

This simulation platform test-bed will allow manufacturer, vendors and suppliers to test their design and their reasoners. It will also show how the system will react with certain faults (fuel leak, electric short circuit) occurring in the system [16].

A. Case study 1 of real accident

In this article the faults that are simulated have been adopted from real accident/incidents. These faults were undetected or miss detected at the system level detection system during the flight, they can be taken as a starting point to see how the VLRS system performs. This requires the system simulation to have certain components which can provide the data to perform higher level reasoning. The following are the accident case studies which have been adopted for this simulation:

*Fault Type:*fuel leak at the entrance of the engine inlet pipe line.

Detection:No fault has been detected at the aircraft system.

1) Details

Flight TS 236 took off from Toronto at 0:52 (UTC) on Friday August 24, 2001 (local time: 8:52 pm (ET) on Thursday August 23, 2001) bound for Lisbon. There were 293 passengers and thirteen crew members on board. The aircraft was an Airbus A330 which was manufactured in March 1999. Leaving the gate in Toronto, the aircraft had 46.9 tons of fuel on board, 4.5 tons more than required by regulation.

At 05:16 UTC, a cockpit warning system chimed and warned of low oil temperature and high oil pressure on engine #2. There was no obvious connection between an oil temperature or pressure problem and a fuel leak. Consequently Captain Piché (who had 16,800 hours flight experience) and First Officer DeJager (pilot who had 4,800 flight hours) suspected they were false warnings and shared that opinion with their maintenance control centre, who advised them to monitor the situation.

At 05:36 UTC, the pilots received a warning of fuel imbalance. Not knowing at this point that they had a fuel leak, they followed a standard procedure to remedy the imbalance by transferring fuel from the left wing tank to the near-empty right wing tank. Unknown to the pilots, the aircraft had developed a fuel leak in a line to the #2 engine. The fuel transfer caused fuel from the left wing tank to be wasted through the leak in the line to the #2 engine. The fractured fuel line, which was leaking at about one gallon per second, caused a higher than normal fuel flow through the fuel-oil heat exchanger (FOHE), which in turn led to a drop in oil temperature and a rise in oil pressure for the #2 engine.

The Portuguese Aviation Accidents Prevention and Investigation Department (GPIAA) investigated the accident along with Canadian and French authorities.

The investigation revealed the cause of the accident was a fuel leak in the #2 engine, caused by an incorrect part installed in the hydraulics system by Air Transat maintenance staff. Air Transat maintenance staff had replaced the engine as part of routine maintenance, using a spare engine, lent by Rolls-Royce, from an older model. This engine did not include a hydraulic pump. Despite the lead mechanic's concerns, Air Transat ordered the use of a part from a similar engine, an adaptation that did not maintain adequate clearance between the hydraulic

lines and the fuel line. This lack of clearance — on the order of millimetres from the intended part — allowed vibration in the hydraulic lines to degrade the fuel line, causing the leak. Air Transat accepted responsibility for the accident and was fined CAD 250,000 by the Canadian government, which as of 2009 was the largest fine in Canadian history.

B. Case Study 2 of real accident

Fault Type: "Fatigue cracking" in a stub pipe within the engine resulted in oil leakage followed by an oil fire in the engine. The fire led to the release of the Intermediate Pressure Turbine (IPT) disc.

Detection: Emergency warnings in the cockpit indicated (engine 2) failure. Pilots were alerted by 54 error messages generated by aircraft systems.

1) Details

Qantas Airline flight 32, Aircraft- Airbus A380, the flight was on route to Sydney Airport via Singapore Changi Airport from London Heathrow Airport on 4th November 2010.

The aircraft engine 2 had an uncontained failure; the shrapnel from this engine had punctured part of the wing and also damaged the fuel system which further caused the problem of leaking fuel and a fuel tank fire. One hydraulic system and the anti-lock brakes were also disabled, which caused engine 1 and engine 4 to go into degraded mode. This meant that the landing flaps were also now damaged.

The failure occurred over Batam Island, Indonesia. After holding to determine aircraft status, the aircraft returned to Changi nearly two hours after take-off. Upon landing, the crew were unable to shut down the (engine 1) which had to be doused by emergency crews 3 hours after landing until flameout. Fuel was leaking from the left wing onto the brakes, which were extremely hot from maximum braking.

An hour after landing the passengers were finally safe to exit the aircraft, there were no injuries to the passengers, crew or people on the ground.

Rolls Royce determined that the direct cause of the oil fire and resulting engine failure was a misaligned counter bore within a stub oil pipe leading to a fatigue fracture.

IV. Architecture of The Simulation Test-Bed

The overview of aircraft vehicle major systems and their architecture is shown in figure 3. In this figure it shows the major systems, which have been modelled in the simulation test-bed. This figure also illustrate basic layout of these systems.

An overview of the experiment architecture for testing and the demonstration of the communication protocol, fault insertion and HM algorithms for the aircraft vehicle is shown in figure 4. The aim of the architecture is to act as a modular test-bed for HM algorithms and data processing architectures from simulation based to embedded hardware implementations [15].

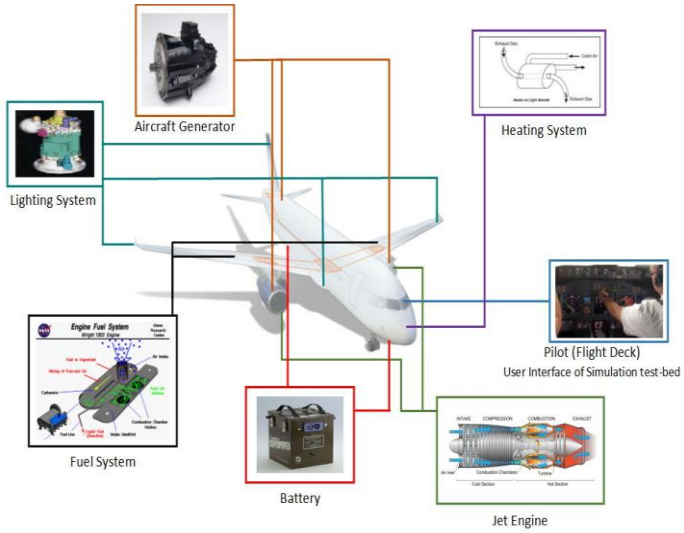


Fig. 3. The architecture of the simulation and physical location of the system in aircraft.

The four main components in the real-time test-bed are the vehicle systems' simulation model, fault insertion unit (FIU), HM data processing and user interfaces (UI). These computation nodes are linked to the Ethernet, and the integrated test bed is enabled by a User Datagram Packet (UDP) based communication between the computers [17],[18].

V. Model of The Simulation of an Aircraft System

In order to evaluate model based diagnosis algorithms, we developed a simulation of the system. The simulation serves as a virtual test bed where we can easily study a large number of fault scenarios to develop our diagnosis models and test our algorithms.

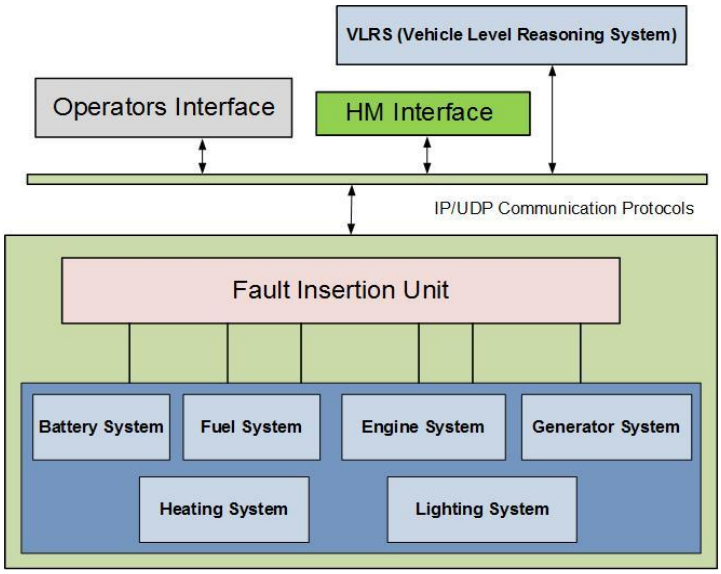


Fig. 4. System Diagram of Simulation test-bed.

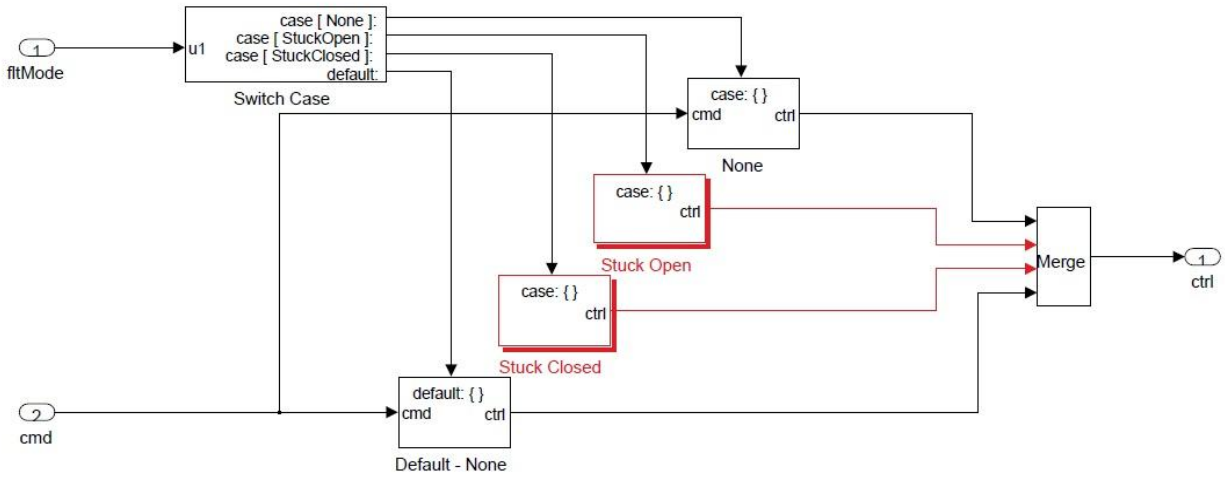


Fig. 5. Fault injection model of the relay component.

An accurate and realistic simulation model will help in migrating diagnosis algorithms to the actual system.

To this end we developed a physics-based simulation of the whole vehicle system in MATLAB/Simulink. We adopted a component based modelling paradigm, where parameterised simulation models of generic components including AC generators, breakers, relays, DC adapters, loads and sensors are available within the SimPowerSystems' component library.

The system model is constructed by instantiating the different components from the component library, specifying their parameters and connecting the components to each other in the appropriate fashion. However, if the required component is not available, we have developed our own by using the Matlab code and performing the task by using mathematical equations.

The simulation test-bed allows for the repeatable injection of faults into the system [19]. All of the fault injections are currently implemented via software. In general, a software fault injection includes one or more of the following: 1) sending commands to the test-bed that were not initiated by the user; 2) blocking commands sent to the test-bed by the user; 3) altering the test-bed sensor data. Because each fault mode is parameterized within the Simulink model, a fault can be inserted either at the beginning of the simulation, or while the simulation is running.

Each component in the simulation model is associated with the fault modes. For example, a relay may become stuck at a particular operating mode. The associated fault injection model of a relay is shown in figure 5, more details of fault insertion are provided in the fault insertion section.

VI. Aircraft Sub-System Modelling

This section will discuss the physics of the each modelled system. The modelling of each sub-system was very important as this simulation test-bed is made to capture the fault progress, how each fault effects the other systems and overall vehicle health.

1) Battery system

Before the engine is started the main source of electrics in an aircraft or in an automotive vehicle are the batteries. The battery also powers up the aircraft systems and brings the aircraft to life before the engine has been started. Once the engines are started the electrical energy to run the system comes from the generators. It also is used to support ground operations such as refuelling and powering the braking system when the airplane is towed. The main battery also provides backup power for critical systems during flight in the unlikely event of a power failure.

In this simulation platform, the battery provides a current before the engine and generators are switched on or in the event there is a need of extra electricity or to store extra electricity. Therefore this simulation platform only monitors the state of the charge of the battery and the charging/discharging rate. However the battery system is extendable.

2) Fuel System

The Fuel system provides the fuel to the engines at the required rate. In the fuel system most of the parts are powered by the electricity. The major fuel system parts are shown in table 2:

TABLE II. Fuel System Parts and Their Quantity

Index	Quantity	Description
1	2	Fuel Tanks (left wing and right wing)
2	2	Hydraulic pump (at left tank and at right tank
3	Several	Hydraulic pipe lines
4	1	Open/close valve
5	1	Pressure Sensor
6	1	Flow Sensor

This fuel system is a small module of the whole simulation platform. The main task of the fuel system simulation is to

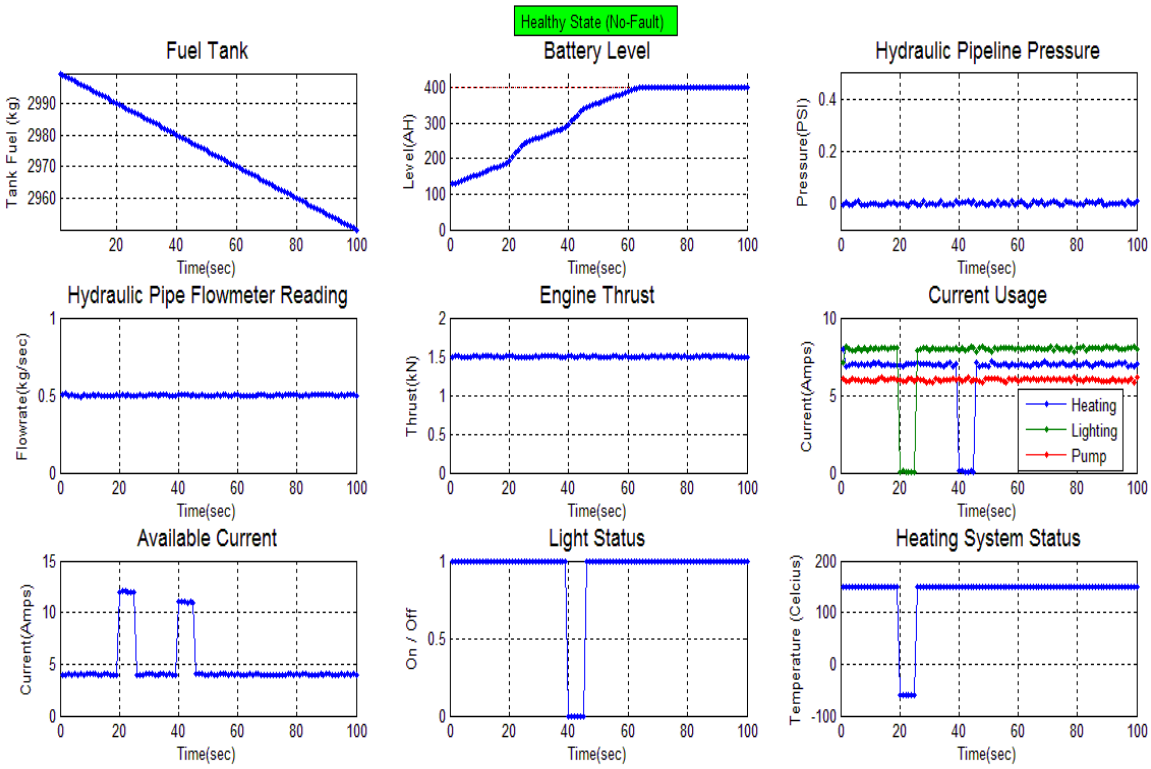


Fig. 6. The Simulation results of a healthy state

provide the data of the fuel system while physically connected with the other systems.This will make the system produce more realistic behaviour data; for example, if the fuel system has any problem and provides less fuel to the engine, then the engine will automatically be affected. However the fuel system has been kept very basic for simple diagnostic tasks. This system provides a flow sensor, pressure sensor, fuel consumption per minute and fuel level at the tank.

In most flows of liquidsand gases at a low Mach number, the density of a fluid can be considered to be constant, regardless of pressure variations in the flow. Therefore, the fluid can be considered to be incompressible and these flows are called incompressible flow. The Bernoulli equation in its original form is valid only for incompressible flow. A common form of Bernoulli's equation is given to calculate the pressure of the pipelines:

$$q = \frac{1}{2}\rho V^2 \quad DynamicPressure \qquad (1)$$

Where:

q : Dynamic pressure $(pascals)$

ρ : Fluid density (kg/m^3)

V : Fluid velocity $(meter/sec)$

3) Engine System

The engine has been simulated as a jet engine, not all the parameters of the jet engine have been simulated at this level as this engine simulation unit is designed for higher level reasoning rather than engine (sub-system) level reasoning. The engine system is physically connected from the fuel and electrical system. The parameters of this engine are fuel intake, air intake, required speed and engine thrust. The mathematical equation has been used in the simulation of the engine unit.

Engine efficiency equation:

The energy efficiency (η) of jet engines installed in vehicles has two main components:

- Propulsive efficiency (η_p): how much of the energy of the jet ends up in the vehicle body rather than being carried away as kinetic energy of the jet.
- Cycle efficiency (η_{ve}): how efficiently the engine can accelerate the jet.

Even though overall energy efficiency η is simply:

$$\eta = \eta_p \eta_{ve} \quad (2)$$

Thrust equation:

The net thrust (F_N) of a turbojet is given by:

$$F_N = (\dot{m}_{air} + \dot{m}_{fuel})v_e - \dot{m}_{air}v \quad (3)$$

Where:

$\dot{m}_{air}$= the mass rate of air flow through the engine

$\dot{m}_{fuel}$= the mass rate of fuel flow entering the engine

v_e= the velocity of the jet (the exhaust plume) is assumed to be less than sonic velocity

v = the velocity of the air intake = the true airspeed of the aircraft

$(\dot{m}_{air} + \dot{m}_{fuel})\, v_e$ = the nozzle gross thrust (F_G)

$\dot{m}_{air} v$ = the ram drag of the intake air

The engine has been modelled closely to the Rolls-Royce RB211, which is part of a family of high-bypass turbofan engines. However the main parameters can get changed by the user to model another jet engine or another type of jet engine.

4) Generator System

The electrical generator system has been modelled to a very basic standard,just as a provider of the electricity at several different speeds. In the civil aircraft industry the generator modules are connected with the engine, each engine has one generator to provide the electricity for the aircraft. Therefore the number of the generator and engine has to be equal in this simulation to make the simulation and functional model of the simulation equal. The generator system monitors the electricity generated by the generator according to the engine and required electricity of the aircraft.

5) Heating System

The aircraft needs heating systems in several places to ensure the safety of the aircraft, for example, a heater at the Pita tubes, turbine blades heaters, front screen heaters etc. Generally these heaters are managed by the heating system. These heaters are electricity powered and provide the required heat at the certain places. The heating unit of the simulation has multidimensional parameters, it consumes the electricity and provides the heat in temperature.

6) Lighting System

The lighting system provides the light to the aircraft in several different places such as the head lamp, tail light, cabin light etc. The lighting system takes the electricity from the main system and the simulation provides the data as to how many bulbs are switched on, how much electricity is being consumed and how much is supposed to be consumed.

The faults can be inserted into all the sub-systems of the whole vehicle system, however, as this simulation platform has simulated very basic sub-systems, not all the parts which are present in a real aircraft system have been available to insert the fault into. Fault insertion modelling is explained in a later stage in this article.

Each component in the simulation model is associated with the fault modes. For example, a relay may become stuck at a particular operating mode. The associated fault injection model of a relay is shown in Figure5.

The Communication Protocol

The communication protocol is a very important part of this project. In order to have a decentralised communication all sub-systems are bound to have information shared between them. Most network based communications is either UDP or TCP based, a comparison of the two is provided below.

TABLE III. TCP PROTOCOL VS UDP PROTOCOL COMPARISON

TCP Protocol VS	UDP Protocol
1. Connection-Oriented	1. Connectionless
2. Reliable (in delivery of messages)	2. Unreliable- No attempt to fragment messages
3. Keep track of order (or sequence)	3. No reassembly, no synchronization and no acknowledgment
4. Use checksums for detecting errors Remote procedures are not idempotent	4. Remote procedures are idempotent
5. Congestion control mechanism is implemented by TCP	5. UDP itself does not avoid congestion, and congestion control measures are implemented at the application level.

The UDP protocol does not support the guaranteed delivery of messages, where on the other hand TCP protocolallows guaranteed message delivery. Therefore the TCP protocol has been used in this simulation.

VII. SIMULATION RESULTS

1) The Initial Results

The simulation test-bed can simulate several different profiles of the system. In the figure below the simulation ran for a 100 seconds without any fault in the system.

This confirms the normal behaviour of the simulation, so the data can be compared with the similar real system.

The light and heating are switched on at the start of the system simulation and the heating switches off at 20sec for 5 sec and the light switches off at 40 sec for 5 sec as shown in the figure9. The General phenomena is visible as the available current increases as the consumption of the current are goes down then the available current are more and other graphs shows the effects as well.

2) The Fuel System Fault Results

The simulation test-bed can simulate several different profiles. In the figure below the simulation has been ran for a 100 seconds and the fault was inserted at 60sec in the system for 10sec.

Figure 7 demonstrates the leakage in the fuel pipe for 10 seconds which shows the behaviour of the engine affected. As the engine didn't get enough fuel the engine thrust level got affected.

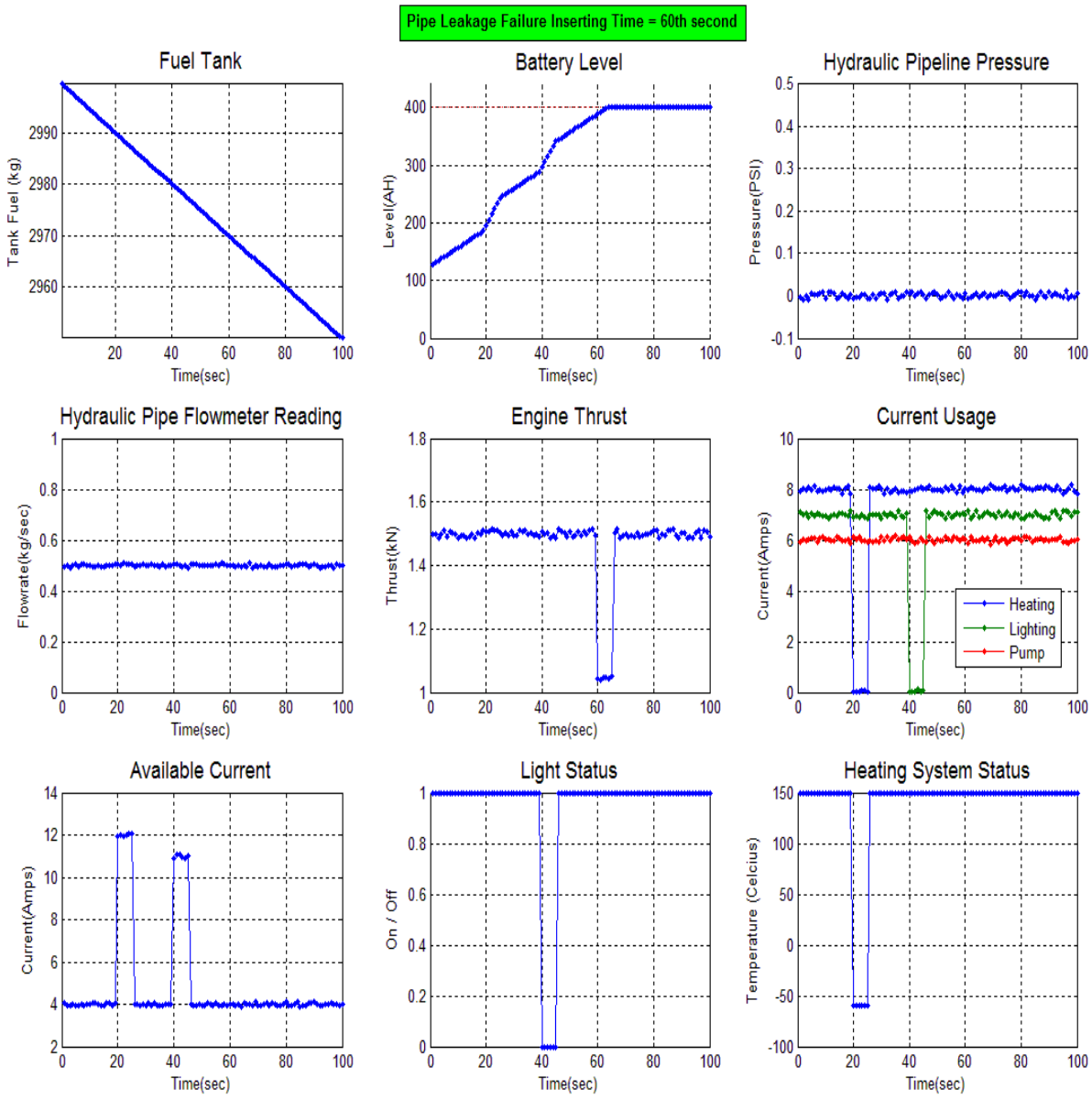


Fig. 7. Fuel Pipe Leakage fault at 60 sec

3) The Heating System Faults Results

The heating system fault has been inserted in the simulation. In the figure 8 the simulation has been ran for a 100 seconds and the fault was inserted at 60sec in the system for 10sec.

Figure 8 demonstrates the short circuit in the heaters of the heating system for 10 seconds which shows the effect on theavailable electric current and it also effected the rest of the electrical system as there wasn't enough current available for other systems to use.

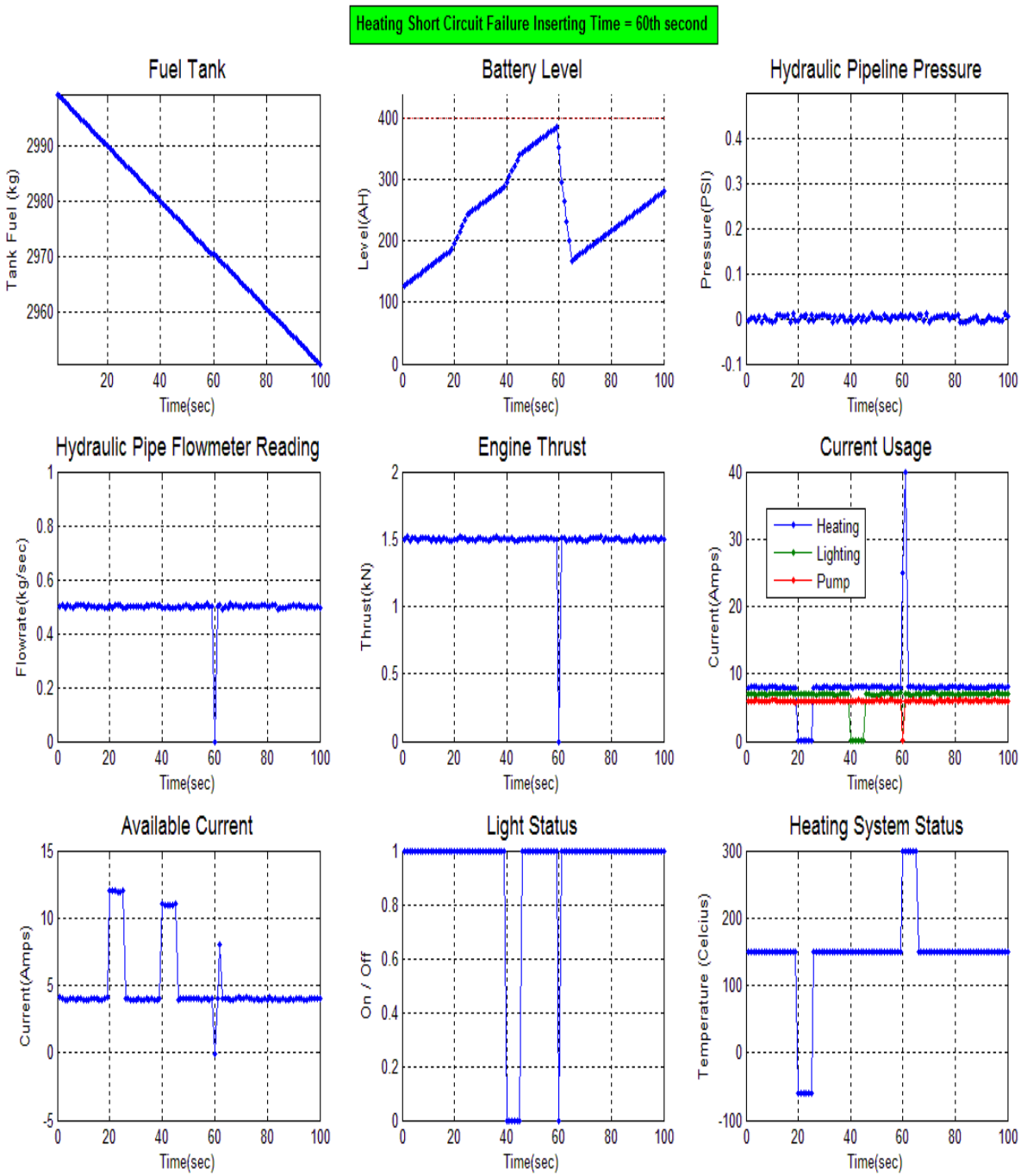


Fig. 8. Fault at heating system due to short circuit.

4) The Lighting System Faults Results

A lighting system fault has been inserted in the simulation result. In the figure below the simulation has been ran for a 100 seconds and the fault was inserted at the 60sec in the system for 10sec.

Figure 9 demonstrates the short circuit in the heaters of the lighting system for 10 seconds which shows the effect on the available electric current and it also effected the rest of the electrical system, there wasn't enough current available for other systems to use.

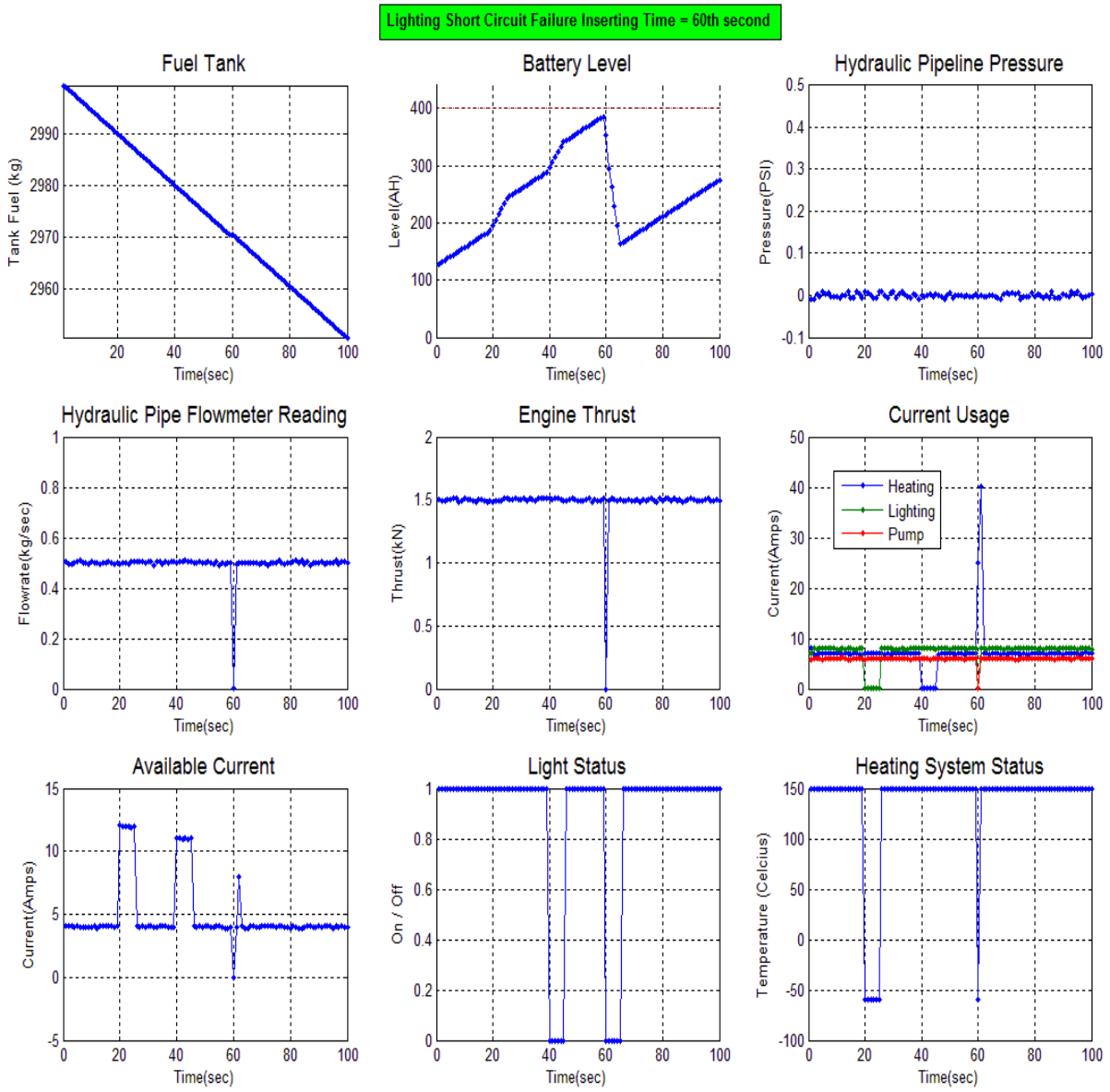


Fig. 9. shows the fault in the lighting system

VIII. Future Works

This simulation test-bed has been designed to produce the data and results on the sub-system level which can be used at VLRS and/or for to apply information exchange between sub-systems. However, in this simulation the VLRS hasn't been implemented. The next stage of this work would be to implement the VLRS and sub-system reasoner and tweak the simulation according to the need of data to reasoners.

IX. Conculsion

The Simulation Test-bed has been designed a as platform to provide data to perform the reasoning at vehicle level. The vehicle level reasoning system provides higher level reasoning results which are achieved by fusing information from the several sub-systems. This simulation platform will provide the data which could be used for the proof of the concept of the efficiency of the vehicle level reasoning system. The simulation platform is very basic compared to the real system however, the data generated from this test-bed would be sufficient enough to provide the health stat and basic fault detection on the vehicle level as well as sub-system level. The next step of this simulation test-bed would be to design the VLRS by utilising the data of this simulation.

X. Acknowledgment

We would like to thank Boeing, BAE Systems and all IVHM centre's industrial partners for their support and guidance in our research. In particular, we would like to thank Kirby Keller from Boeing, and Antony Waldock from BAE

systems, for their collaboration and all the feedback provided to us on our research direction.

REFERENCES

[1] F. Khan , I. Jennions and T. Screenuch, "Integrated Issues for Vehicle Level Distributed Diagnostic Reasoners," SAE Technical Paper, 2013.

[2] A. Del Amo, K. Keller and K. Swearingen, "General reasoning system for health management," in Fuzzy Information Processing Society, IEEE, 2005.

[3] J. d. Kleer and J. Kurien, "Fundamentals of model-based diagnosis," in In proceedings of IFAC Safeprocess, Washington, USA, 2003.

[4] J. Luo and Z. Su, "Design and Implementation of Intelligent Diagnostic System Based on AI-ESTATE," Fourth International Conference on Information and Computing , pp. 237-240, 2011.

[5] T. Sreenuch , A. Tsourdos and I. K. Jennions, "Distributed embedded condition monitoring based OSA-CBM Startdard & interfaces," pp. 238-246, 2013.

[6] B. Chidambaram, D. Gilbertson and K. Keller, "Condition based monitoring of an electro-hydraulic system using open software architectures," in IEEE Aerospace Conference , 2005.

[7] N. V. Bedina, "Fault Simulation and Diagnostics in Generating System for Aircraft," Electronic Devices and Materials , pp. 141-143, 2007.

[8] A. Bajwa and A. Sweet, "The Livingstone model of a main propulsion system," in IEEE Aerospace Conference, 2003.

[9] W. T. Thomson and R. J. Gilmore, "Motor current signature analysis to detect faults in induction motor drives-Fundamentals, data interpretation, and Industrial case histories," in Proceeding of the Thirty-second Turbo machinery Symposium , 2003.

[10] W. R. Finley, M. M. Hodowanec and W. G. Holter, "An Analytical Approach to Solving Motor Vibration Problems," in Petroleum and Chemical Industry Conference, Industry Applications Society 46th Annual , San Diego, 1999.

[11] D. Followell, D. Gilbertson and K. Kelly, "Implications of an open system approach to vehicle health management," in IEEE Aerospace Conference , 2004.

[12] R. Isermann, "Model-based fault-detection and diagnosis - status and applications," Annual Reviews in Control Vol.29, pp. 71-85, 2005.

[13] A. Srivastava , R. Mah and C. Meyer, Automated Detection Diagnosis, and Prognosis to enable mitigation of adverse Events during flight, NASA, 2009.

[14] S. Liao, "Expert system methodologies and application- a decade review from 1995 to 2004," Expert Systems with Applications, pp. 93-103, 2005.

[15] I. Jennions, Integrate Venhicle Health Management, Perspectives on an Emerging Field, Warrendale, Pennsylvania, USA: SAE International, 2011.

[16] D. Gorinevsky, S. Boyd and S. Poll, "Estimation of faults in dc electrical power system," in American Control Conference, St. Louis, 2009.

[17] "MathWorks," Real-Time Windows TargetTM 3: User's Guide , Natick, MA, 2010.

[18] A. J. Giarla , "A declaration based logic calculator AUTOTESTCON Proceedings," in Systems Readiness Technology Conference, 2001.

[19] A. Saxena, K. Goebel, D. Simon and N. Eklund, "Damage propogation modeling for aircraft engine run-to-failure simulation," Prognostics and Health Management, pp. 6-9, 2008.

Selection of Touch Gestures for Children's Applications: Repeated Experiment to Increase Reliability

Nor Azah Abdul Aziz, Nur Syuhada Mat Sin
Creative Multimedia Department
Faculty of Art, Computing and Creative Industry
Sultan Idris Education University,
Malaysia

Firat Batmaz, Roger Stone and Paul Wai Hing Chung
Computer Science Department
Loughborough University
Leicestershire, United Kingdom
F.Batmaz, R.G.Stone,

***Abstract*— This paper discusses the selection of touch gestures for children's applications. This research investigates the gestures that children aged between 2 to 4 years old can manage on the iPad device. Two experiments were conducted for this research. The first experiment was carried out in United Kingdom. The second experiment was carried out in Malaysia. The two similar experiments were carried out to increase the reliability and refine the result. This study shows that children aged 4 years have no problem using the 7 common gestures found in iPad applications. Some children aged 3 years have problem with two of the gestures. A high percentage of children aged 2 years struggled with the free rotate, drag & drop, pinch and spread gestures. This paper also discusses the Additional Criteria for the use of Gestures, Interface Design Components and Research on Children using iPad and Applications.**

Keywords— Children; Gesture; Applications (Apps)

I. INTRODUCTION

Gestures are defined by [1] as a powerful feature of human expression, either alone or as a means for augmenting spoken language. This paper is focusing on our experiment in Malaysia which is the latest phase of our study which started in 2012. In this research four education and games category apps for the iPad were selected from the Apple store and seven gestures were chosen for study as used in our previous research for young children aged 2 to 4 years in United Kingdom [2]. This paper reports the result of an experiment carried out on children in Malaysia and highlights the differences and similarities compared to the previous similar experiment carried out in the United Kingdom.

This paper is divided into the following sections: Literature Review, Experiment Set Up, Results & Discussion, Conclusions.

II. LITERATURE REVIEW

Play is not only an enjoyable and spontaneous activity of young children but it also contributes significantly to children's psychological development [3]. Children are always being curious and want to explore new things in their life. It is this curiosity that makes the touch screen technology so popular. With just a touch of a finger, children can interact with a smart phone or tablet [4]. When teachers or parents use technology that children are comfortable with, they may be encouraged to learn through playing.

The result from the research which compares the use of tablet internationally in 2012 [5] shows that the use of tablets is not high in many countries with 18% in Egypt and Chile, followed by 7% in Indonesia and 5 to 7% in Japan and India. The use of tablets by children may be hindered by the high purchase price. Wider use of tablet is expected due to the price drop every year. The research [5] compares the use of tablet in general but not specifically in school. The use of iPad tablet in large numbers in Alberta classrooms was recorded in the Alberta Summary Report [6] with 147 participants, representing 25 school authorities. They noticed that the use of iPads has increased student and teacher engagement, improved the capacity to meet a wide variety of learning needs and provided more ways for students to demonstrate their understanding. Even though tablets, especially iPad, are not widely used among children generally [5], they are used in classrooms like Alberta schools where government funding was available.

The study [7] is about the cultural and economic differences on the use of mobile phones and computers by children in general and was done with native Dutch and Immigrant children aged 4 to 7 years. There are no significant differences found between the attitudes of Dutch and immigrant children in using computers. The children from a lower socio-economic neighborhood had more positive attitudes towards computers and used computers slightly more often than middle class children [7]. The findings [7] showed that culture does not influence children in using mobile phone and computers.

Research is necessary to understand and enable the real benefits of these increasingly popular technologies [8]. Providing an experiment on how children use gesture has great potential to provide design guidance and positively influence children's digital experiences with these new forms of technology [4, 8].

The study by [9] of using iPad as a learning tool for children between 8 and 12 years old and their teachers related to the design activities and the use of iPad application. The children evaluated the use of iPad as truly successful.

However the children's criteria of success were how fun and enjoyable it was to use the iPad, and the ability to work in teams and try something new. Meanwhile their teachers found that the use of iPad did not improve the children's learning outcome [9].

According to [10] the children felt that the technology supported them appropriately when it offered them control, no matter how small, of their physical interactions. The developing of physical coordination skills and physical sizes of young children place restrictions on using technology and this is the challenge for designers when developing physical interfaces for young children.

Like the Alberta Schools which used iPad in school [6], the study by Michael Cohen Group's also discussed children's ability in using iPad in general. However, the study did not focus on the use of touch screen gestures in detail [11].

Recent studies that related to our study are more on the use of touch screen and gestures in general by young children [2]. The most closely related study was done by Sesame Street. The creators of Sesame Street discuss in the short section on gestures that they have identified the most and the least intuitive gestures for preschool aged children. They have found tap, draw/move finger, swipe, drag and slide to be the most intuitive gestures. Pinch, tilt/shake, multi-touch, flick/fling and double tap are the least intuitive gestures [12]. However they do not draw any distinction among children by age.

To the best of our knowledge, no research has investigated and compared the gestures among children by age.

This paper is focusing on our experiment in Malaysia which is the latest phase of our study which started in 2012. The first phase of this study was to identify the common gestures used in children's applications [4], select the appropriate applications (apps) for the experiment and carry out the pilot study [13] and experiment in the United Kingdom [2].

The common gestures used in children's applications were found to be Tap, Drag/Slide, Free Rotate, Drag & Drop, Pinch, Spread and Flick. These gestures were found in 100 children's applications from the Apple store [4].

Following on from the pilot study which investigated children aged 2 to 12 years (3 children for each group) in using 7 gestures in United Kingdom, we conducted an experiment with 37 children aged 2 to 4 years in the United Kingdom. Seven types of gestures were chosen on iPad applications. The results from our study in United Kingdom showed that all gestures can be used by children at aged 4 years and children aged 2 to 3 years have problems using certain gestures. Therefore, this study will use the same experiment design for aged 2 to 4 years with the same seven gestures with children of a different culture in Malaysia in order to increase reliability and allow us to refine the result.

III. Experiment Set Up

We use the same experiment set up as our previous experiment in United Kingdom. This repeated experiment is to increase reliability and allow us to refine the result. We believe that the repetition of experiment reduces the possibility of errors and also verifies the accuracy of the previous findings.

In terms of experiment methodology, according to [14] researching with children is different from researching with adults. The researcher has to establish a friendly relationship with children. The researcher has to interact with them in the most trusted way possible without having any explicit authority role. Therefore more time is needed to interact with children and gain their trust. In this experiment we have spent more time with each individual child in the experiment Malaysia's experiment compared with the experiment in the United Kingdom.

The four selected applications are: Montessori Crosswords (English versions), AlphaBaby Free (English versions), Toca Hair Salon and Toca Kitchen Monsters.

An interface should use language and concepts that the user is familiar with [15]. The method of selection of the 4 applications in our research included a check that the applications used language and concepts that children are familiar with. The Malay language is the first language for children in Malaysia but they also learn English as a second language from preschool.

Toca Hair Salon and Toca Kitchen Monsters did not use language as a medium for user interaction. Both apps provide a gestural interface for young children. Hair Salon provides gestures for combing, cutting, spraying, and coloring hair. Toca Kitchen Monsters used the kitchen theme and allows the children to use gestures to choose food, cook and feed the monsters.

The Alpha Baby application provides young children with a gestural interface for learning basic alphabet, numbers and shape using the English language. The Montessori Crosswords app maybe the most challenging one for young children to interact with because they have to drag letters to form a word based on images showed on the screen. For young children who do not know how to read, the teacher and researcher will show them the correct letters they have to choose in order for them to perform the gestures. The teacher and researcher have to bear in mind that the experiment is to evaluate the children's capability to use different gestures and not their spelling.

Forty children from National Children Development Research Centre (NCDRC), Sultan Idris Education University, Malaysia participated in this study. The children were aged:

1) *2 years (10 children)*
2) *3 years (14 children)*
3) *4 years (16 children)*

The children used an iPad one at a time in a comfortable environment. The researcher together with the teacher guided the child to play with each application. Based on past experience, more time was given to children aged 2 and 3 years to play and familiarize themselves with the applications and gestures. The children were given the opportunity to use the same gesture 3 to 5 times before their gestures were being

recorded. A digital video camera was used to record the gestures made by each child.

IV. Results and Discussion

This section is divided into the following sub-sections: Gestures that can be used by Children, Additional Criteria for the use of Gestures, Interface Design Components and Research on Children using iPad and Applications.

A. Gestures that can be used by Children

Seven common gestures were selected for this experiment: tap, drag/slide, free rotate, drag & drop, pinch, spread and flick. Table I summarizes the results of the analyzed video recorded during experiment with 40 children aged 2 to 4 years, in Malaysia and 37 children aged 2 to 4 years in United Kingdom .

TABLE I. Gestures That Can Be Used By Children Aged 2 To 4 Years

Gestures	Age 2		Age 3		Age 4	
	M	UK	M	UK	M	UK
Tap	100%	100%	100%	100%	All 100%	
Drag/Slide	100%	100%	100%	100%		
Free Rotate	40%	55%	100%	91%		
Drag & Drop	30%	36%	100%	100%		
Pinch	30%	55%	71%	82%		
Spread	10%	11%	64%	36%		
Flick	80%	36%	100%	73%		

M=Malaysia UK =United Kingdom

The table shows the list of gestures in the first column and the following columns show the percentages of children aged 2 to 3 years who can use the gestures. We exclude the details result for children aged 4 years old in Table I because all of them could use all 7 gestures successfully in both the Malaysia and United Kingdom experiments.

This research consistently considers that a percentage lower than 70% indicates that the children are struggling in using gestures and a higher percentage indicates they are successful in using the gestures.

Table I shows that all children from aged 2 to 4 years can use the tap and drag/slide gesture. The gestures are easy and natural as this is what children see and do in their real life. The research [12] also found that tap is the most intuitive and foundational touch interaction for children.

The free rotate gesture requires the children to twist their fingers. It is shown that only 40% children from aged 2 years from Malaysia and 55% from United Kingdom can use the free rotate gesture. This indicates that children aged 2 years old are struggling to execute the free rotate gesture. The observation shows that children aged 2 years did not have the capability to twist their fingers easily such as rotating a letter in the Alpha Baby application. All children aged 3 and 4 years can use free rotate gestures. This study confirms that children aged 3 and 4 years have no problem in using this gesture.

The drag & drop gesture requires the children to press and move their finger without losing contact with the surface. Only 30% of children aged 2 years from Malaysia and 36% from United Kingdom can use drag & drop gesture. This indicates that children aged 2 years old are struggling to execute the drag & drop gesture. The reason may be either that they do not understand how to do the drag & drop gesture or their motor skill is not yet fully developed or both. The children were required to use the drag and drop gesture in the Toca Kitchen Monsters, Toca Hair Salon and Montessori Crossword applications. All children aged 3 and 4 years were successful in using the drag & drop gesture.

The pinch gesture requires the children to use two fingers and bring them closer on the surface. The result also shows that only 30% of the children aged 2 years from Malaysia and 55% from United Kingdom can use the pinch gesture. This indicates that children aged 2 years are also struggling to execute the pinch gesture. For example children aged 2 years were struggling to pinch the shape to make it smaller in size in the Alpha Baby application. Perhaps they lack the capability to perform pinch gesture (motor skill) or do not understand how to do the gesture (cognitive level). Meanwhile, 71% children from Malaysia and 82% from United Kingdom aged 3 years and all children aged 4 years can use the pinch gesture.

The spread gesture requires the children to touch the surface with two or more fingers and move them apart. With the spread gesture, 10% of the children aged 2 years from Malaysia and 11% from United Kingdom can use it. This indicates that children aged 2 are struggling to execute the spread gesture. Like free rotate, drag & drop and pinch, it is assumed the reason for the struggle maybe either that they do not understand how to do the spread gesture or their motor skill is not yet fully developed or both. 64% of the children aged 3 years from Malaysia and 36% from United Kingdom can use the spread gesture. This also indicates that children aged 3 years are struggling using spread gesture even though the table shows the percentage were increased. The increased percentage maybe because they were given an opportunity to use every gesture 3 to 5 times before the researcher record their fingers movement.

The flick gesture requires the children to use their finger to brush the surface. With the flick gesture, surprisingly 80% of the children aged 2 years from Malaysia can use it even though only 36% children aged 2 years from United Kingdom can use the gesture. The increased percentage maybe because they were given an opportunity to use every gesture 3 to 5 times before the researcher recorded their finger movement. This may suggest that 2 years old children could learn some of the gestures with practice. This indicates that Malay children aged 2 years have no problem with flick gestures as well as children aged 3 and 4 years.

The results in Table I show that Malaysia's experiment seem to be consistent with the results of the experiment in the United Kingdom except for a surprising percentage increase in the flick gesture for 2 year old and the spread gesture for 3 year old children.

These results also confirm the full capabilities of children aged 4 years in using all seven gestures. Meanwhile children aged 3 years also struggling using spread gesture. Attention should be given to children aged 2 years old who are struggling to use more than half of the seven gestures. Therefore developers should be careful with the application design for children aged 2 and 3 years. The selection of gestures needs to be appropriate to the children's age for application design and this experiment shows a strong relationship between age and the gestures that can be used.

B. Additional Criteria for the use of Gesture

Four additional criteria for the use of gestures are identified in the previous UK study.

1) *Unique gesture or 'one gesture - one task',*
2) *simultaneous gestures,*
3) *consistent gesture and*
4) *natural gesture.*

Unique gesture is one gesture implemented to achieve any given task on a particular component. Simultaneous gestures can refer to the set of all gestures used in all tasks on all components or the ability to apply a gesture to more than one component at once. The requirement for consistency in gestures is the use similar gesture for a similar task on other components on other screens. The natural gesture is where the use of the gesture is consistent with its use in the real world [2]. The observation and the analysis of the records for the Malaysian experiments confirm the conclusions about the additional criteria previously as previously identified.

Children's applications come with components which are touchable such as shapes, letters, numbers or objects. As in our previous experiment, our observation shows that simultaneous gestures have no meaning to young children aged two and three years. 90% of the children aged 2 to 3 years only use one unique gesture for each task even though they are taught to use other gestures by the researcher or their teacher. For example the Alpha Baby application was designed with 6 gestures on the same screen. Children could use tap to allow the component (number/alphabet/shape) to appear on the screen and use other gestures such as drag/slide, pinch, spread, flick and free rotate to move, resize, animate and rotate the components. Our observation shows that the children keep tapping a letter or digit on Alpha Baby application instead of using other gestures. Therefore the designer could consider using just the tap gesture or another unique gesture for Alpha Baby application.

The inconsistent gestures make children confused. For example, 80% children aged 2 to 4 years in Malaysia and the same percentage of children aged 2 to 4 years in United Kingdom were using the drag/slide or tap gesture even though they have to do drag & drop gesture in the final screen in Toca Hair Salon. This is because every screen in Toca hair Salon used the same gesture except the final screen. The children looked confused at the beginning until the researcher asked them to use the correct gesture. The children are able to use the application smoothly if the application's designers use consistent gestures for the whole application.

Our observation in Malaysia's experiment also shows that 50% of children always use natural gestures at the beginning of their interaction with applications. For example the children try to pick the food from refrigerator (using all fingers like pinch gesture) in Toca Kitchen Monsters application instead of dragging and dropping the food into the monsters mouth. The designer may use pinch gesture instead of drag and drop to select food from the refrigerator for the Toca Kitchen Monsters application.

Our previous research finding also shows that young children aged 2 to 4 years old need a natural interface design for touch screen application. Little children playing or using the application on the touch screen instinctively followed the way they have done it in the real world such as picking object, feeding pet and combing hair by using all fingers like the pinch gesture [2].

C. Interface Design Components

This section discusses the interface design components which are touchable, e.g. a shape, a letter, a number or any object in a children's application [2].

The observation for the Malaysian experiment is also consistent with the previous experiment that too many components for every page will keep children busy sorting the components (numbers/letters/objects) rather than answering the question given on that page and the children also found it difficult and got confused when using gestures on crowded interface design such as in the Montessori Crossword and the Alpha Baby application. 50% of children aged 3 to 4 years were busy sorting the components meanwhile 100% children aged 2 years were only tapping and staring at the components on the screen. The application designer should consider simple design such as one component for each screen, consistency of the components and their arrangement throughout the whole application such as using one type of component for each screen and place it in the middle of the screen.

40% of the children aged 3 to 4 years also relate the image or item they see on the application to what they usually see in the real world. For example the children try to rotate the water tap to wash the cartoon character's hair in the Toca Hair Salon application instead of using a tap gesture. The children also try to use inactive images such as butter together with other food in Toca Kitchen Monsters. The selection of image and gesture must be consistent with its use in the real world as mentioned in Section IV-B because children will use natural gestures associated with the image.

Observation also shows that 20% of the children aged 2 years old did not want to play with certain cartoon characters in Toca Kitchen Monsters and Toca Hair Salon when the researcher showed the applications. They kept crying and ignored the application throughout the experiment. Therefore, before using any image, especially a cartoon character, the application's designers have to test it on young children to ensure the character does not scare them.

When designing an interface, a designer is encouraged to reduce the number of components for each page, select appropriate images/characters for young children and arrange

the components in a consistent and simple way for the whole applications.

D. Research on Children using iPad and Applications

Observation also shows that children aged 2 years old need more time to familiarize themselves with the applications. They also need more time to interact in a friendly mode with the researcher who was an outsider to them. Male children aged 2 years old are friendlier than female but in terms of understanding and following the instructions, the female children are better. Overall observation shows children aged 2 years old were more likely to feel uncomfortable, become distracted or cry. For much of the time the teachers and researchers had to hold their hand and show them how to use the gestures correctly. They were also uncontrollable and used their fingers on the screen without any purpose or understanding what they were doing.

Children aged 3 years old were also eager to use the iPad and the applications especially the male children. The researcher had to guide them to use all seven gestures but they required less time to understand how to use the gestures and interact with the application in comparison to children aged 2 years old. Children aged 3 years old were also easily influenced by their friends. They were eager to play with the application when they saw their friends interacting with it.

Children aged 4 years old are highly motivated or eager to explore all the applications. They also understand how to use the gestures and applications easily. This observation confirmed that children aged 4 years have no problems using all the seven gestures and interacting with all four applications. Therefore the designer may use all seven gestures for children aged 4 years.

Overall observation shows that the children's favourite application is the Toca Kitchen Monsters and more than half of the children did not like to play or interact with Montessori Alphabet. The Toca Kitchen Monsters application is popular among children and this may be because of the use of consistent gesture (drag & drop), and because feeding the monsters is related to what the children do in the real world (children eat or their mother feeds them every day).

V. Conclusion

An overall observation indicates that the children's age, components criteria and interface design components influence the children's capability of using gestures in every screen.

The results of this experiment are consistent with our previous experiment which was done in United Kingdom except for the flick gesture for children aged 2 years old and the spread gesture for children aged 3 years as shown in Table II.

TABLE II. Gestures That Can Be Used By Young Children Aged 2 To 4 Years

Malaysia		
Age 2	**Age 3**	**Age 4**
Tap Drag/Slide Flick	Tap Drag/Slide Drag & Drop Free Rotate Pinch Flick	Tap Drag/Slide Drag & Drop Free Rotate Pinch Flick Spread

Children aged 2 years old successfully used the flick gesture and the percentage of children aged 3 years old also increased in using the spread gesture in Malaysia's experiment. This may be because more time was allocated for them to play with the gesture and application compared to our experiment in United Kingdom. Our observation shows that there are no significant differences in the understanding and use of gestures between the children in Malaysia and United Kingdom arising from the differences in culture. The result for this repeated experiment confirmed that gestures that can be used by children aged 2 to 4 years old. This result suggest that application designer can use tap, drag/slide and flick gestures in children's applications aged 2 years and add drag & drop gesture for children aged 3 years and all seven gestures for children aged 4 years and above.

The application designer also needs to consider the children's response to other gestures criteria, interface design components and how the children interact with applications and iPad. The way that an application designer chooses and arranges the gestures must be appropriate. The application designer should use a unique gesture for a given task, consistent gestures across different screens and natural gestures for the whole application. When designing an interface, a designer is encouraged to reduce the number of components for each page, select appropriate images/characters for young children and arrange the components in a consistent and simple way for the whole applications.

Acknowledgment

Thanks to the children, parents and teachers from National Children Development Research Centre (NCDRC), Sultan Idris Education University, Malaysia who agreed to take part in this research.

References

[1] S. Constantine, "Natural interaction through gesture recognition and head and body tracking", HCI newsletter Issue Number 58, March 2013.

[2] A. A. Nor Azah, B. Firat, S. Roger and W. H. C. Paul, "Selection of Touch Gestures for Children's Applications", Science and Information Conference (SAI) 2013, October 7-9, 2013 , London.

[3] V. Irina, H. Pauline and L. Pauline, "Child's Play: Computer Games,Theories of Play and Children's Development", In Proceedings of CRPIT '03, Australian Computer Society, Inc., vol 34, 99-106, 2003.

[4] A. A. Nor Azah, "Touch Screen Application (iPad): The most used Gestures for children's applications, *Aplikasi Skrin Sesentuh iPad: Gerakan Jari yang selalu digunakan untuk Kanak-Kanak*", NCDRC, Journal, in press.

[5] GSM Association and the Mobile, Society Research Institute within NTT DOCOMO Inc. Japan, "Children's use of mobile phones-An International Comparison 2012", 2013.

[6] Alberta Government of Alberta, "iPads: What are we learning?", Summary Report of Provincial Data Gathering Day, October 3, 2011.

[7] M. K. Susan and J. Voogt, "Technology and young children: How 4–7 year olds perceive their own use of computers", Computers in Human Behavior, 2010, Vol.26(4), pp.656-664 SciVerse ScienceDirect Journal.

[8] A. N. Alissa, "Knowledge gaps in hands-on tangible interaction research", In Proceedings of International Conference of Multimodal Interaction (ICMI '12), ACM Press, Santa Monica, CA, USA, Oct 22-26, 2012.

[9] Alma, G. And G. Andrea, "Tweens with the iPad Classroom – Cool but not Really Helpful?", International Conference on e-Learning and e-Technologies in Education (ICEEE), 2012.

[10] M. Jaime, D. Allison, C. Gene, F. Allison and L. Mona, "Tools for Children to Create Physical Interactive StoryRooms", ACM Computers in Entertainment, Volume 2, Number 1, Article 3, January 2004.

[11] Michael Cohen Group LLC. Young Children, Apps & iPad. *Michael Cohen Group LLC Report*, 2011.

[12] Sesame Street, "Best Practise: Designing touch tablets experiences for preschoolers", 2012, http://www.sesameworkshop.org/assets/1191/src/Best%20Practices%20Document%2011-26-12.pdf.

[13] A. A. Nor Azah, "Children's Interaction with Tablet Applications: Gestures and Interface Design", International Journal of Computer and Information Technology, Vol. 02, Issue 03, 447-450, 2013.

[14] G. A. Fine and S. L. Sandstrom. "Knowing Children: Participant Observation with Minors", Monograph on the techniques and ethical issues of research with preschoolers through adolescents, Newbury Park: Sage Publications, 1988.

[15] G. Heléne and K. Paula, "Ten Design Lessons from the Literature on Child Development and Children's Use of Technology", IDC 2009, 52-60, ACM 2009.

A web based Publish-Subscribe framework for Mobile Computing

Cosmina Ivan
Department of Computer Science
Technical University of Cluj Napoca
Cluj, Romania

***Abstract*—The growing popularity of mobile devices is permanently changing the Internet user's computing experience. Smartphones and tablets begin to replace the desktop as the primary means of interacting with various information technology and web resources. While mobile devices facilitate in consuming web resources in the form of web services, the growing demand for consuming services on mobile device is introducing a complex ecosystem in the mobile environment. This research addresses the communication challenges involved in mobile distributed networks and proposes an *event-driven communication* approach for information dissemination. This research investigates different communication techniques such as polling, long-polling and server-side push as client-server interaction mechanisms and the latest web technologies standard *WebSocket,* as communication protocol within a Publish/Subscribe paradigm. Finally, this paper introduces and evaluates the proposed framework, that is a hybrid approach of WebSocket and event-based publish/subscribe for operating in mobile environments.**

Keywords—mobile computing; Websockets; publish-subscribe; REST

I. Introduction

In recent years, the growth of mobile devices such as smartphone and tablets has led to an extensive use of mobile applications in almost every sector of our life. The Gartner research [1] forecast 2011 states, that the download of mobile apps worldwide had increased by 117 percent from 2010 to 2011 and forecasts an astounding 185 billion downloads from mobile app store by 2014, since the first launch in 2008. The capabilities of these devices in doing more than just making calls as well as sending and receiving text messages has increased the demand for mobile applications in the enterprise, as it becomes possible for enterprises to extend their services to the fingertips of numerous consumers.

Generally, these mobile applications consume data as Web services from a remote server- based architecture, which is the backbone of most information systems. Today's information society is built upon collaborative platforms which gathers and shares information across distributed networks, so the backbone of these information systems consists of multiple disparate system applications. The growing demand of consumers in accessing services is causing these systems to expand and some of these services can be hosted in the cloud computing environment, in order to ensure availability, reliability and scalability in service consumption. Cloud computing is the era where IT services are outsourced from providers over the internet on pay-according-to-use policy [2]. With the growing demand of consumer web services and the expansion of systems that forms a gigantic distributed heterogeneous infrastructure, there is an acute need for frameworks that can reliably operate in the mobile environment.

The remainder of the paper is organized as follows. Section 2 reviews some of the key points that this study explored and the existing research works within the identified problem domain. Section 3 presents the proposed framework design in addressing the research goals and challenges. Section 4 describes the implementation details of the architecture followed by the experiments designed to verify the framework in accordance with the research goals. Finally, section 5 concludes the thesis with the contributions of this research.

II. Problem Specification and Literature Review

While distributing the system applications provides more flexibility and scalability, it often results into a growing system complexity during services consumption in a mobile environment. One of the major challenges in today's enterprise solution is to ensure integration among these disparate and distributed system applications which are often connected to legacy systems. In addition to that, mobile devices are becoming an integral part of the growing digital ecosystem and the primary means of accessing IT services. The major challenges while disseminating data over a wireless connection in a mobile environment are as follows: unreliable network connection**,** higher degree of network latency, limited network bandwidth.

This introduces more challenges to the system when synchronizing the information flow between mobile clients and the distributed system backend.

A. Problem specification

In addressing the above mentioned challenges in mobile digital ecosystems, this research looks into *developing a framework* for disseminating data over wireless networks and *proposes an architecture* that allows system components to independently propagate data (i.e. resource updates) and as they propagate, the *eventual consistency technique* is employed to synchronize the data . In this regard, this paper looks into the *Pub/Sub pattern* as a mechanism for propagating data close to real-time, moreover, the emergence Web 2.0 has greatly embraced the Restful web services [3] due to its web compliant API and lightweight solution for

resource's state management. Therefore, the proposed framework is a hybrid of REST-based and event-based Pub/Sub that deploys a combination of various client-server interaction modes, such as polling, long- polling and server-side pushing.

The main research goal in proposing such a framework for mobile devices is to integrate REST web services within Pub/Sub domain. In this respect, the research will look into different *Rest patterns* in disseminating data, chose the most suitable for an event-based Pub/Sub system and address the above mentioned challenges in wireless network. The secondary research goal is to reduce network latency, bandwidth usage and also synchronizing resource's state in the face in intermittent connection loss, in terms of the proposed implementation.

The remainder of the paper is organized as follows. Section 2 reviews some of the key points that this study explored and the existing research works within the identified problem domain. Section 3 presents the proposed framework design in addressing the research goals and challenges. Section 4 describes the implementation details of the architecture followed by the experiments in section 5 designed to verify the framework in accordance with the research goals. Finally, section 6 concludes the thesis with the contributions of this research.

B. Literature review

A communication model that helps in dealing with the information dissemination in a large scale mobile network is Pub/Sub paradigm [4]. In this Pub/Sub architecture, information providers as publishers disseminate information in the form of events and information consumers ,as subscribers register for events of their own interests. There can be an event broker acting as a middleware which helps in dispatching events to the respective subscribers.

Communication in Pub/Sub is inherently asynchronous and transparent in nature as both entities (information provider and subscriber) operate asynchronously through a dispatcher and disseminate state changes to all interested subscribers through one operation. In the basic model of a Pub/Sub system, both providers and subscribers are connected through a set of groups or channels through which subscribers are notified for the events of their interest. Upon receiving event notification, the publisher dispatches the event to the respective subscribers.

As subscribers are not interested in all the events that are published by the providers, there are various ways that the subscriber can specify interest for a specific event. These variations have led to different subscription models that are currently seen in Pub/Sub system environments. The most important subscription models are topic and content based schemas.

One of the first generation subscription schemes is the topic-based scheme. In this scheme, subscribers register for notification based on the topic or subject of the events corresponding to a particular group ,or a set of groups also known as a logical channel [5]. Users subscribed to a channel(s) will receive all published events of that channel. The topic-based scheme has been proposed as a solution in many industrial Pub/Sub environments, one of the most mentioned systems is CORBA notification service and DDS from OMG group, also among others, and TIB/RV, SCRIBE and Bayeux are some of the systems that implement topic-based scheme **[**5**]**.

The Pub/Sub paradigm is better understood in the domain of a *messaging system* and also known in the domain as Pub/Sub messaging system, and has the capability of managing messages in a similar way that a persistent database is managed by a database system. Messages are coordinated and integrated among the software components as software applications changes over time, and are transferred from one machine to another over the unreliable wireless network.

A more flexible but also complex paradigm in the Pub/Sub scheme is content-based subscription. It provides more flexibility to the subscriber by providing more control in subscribing an event based on the actual content of the event. It allows subscriber to impose set of constraints in the form of condition in forming a query on an event notification (also known as *filter*). Creating a notification using a filter provides subscribers with a more sophisticated way for subscribing events. However, this higher expressive capability in defining subscription on the other hand, can be an added challenge in implementing such a scheme, since matching publisher's events with subscriber become more complicated and the resource consumption becomes higher, inappropriate to our goals [5]. There are several examples of systems that implement content- based subscription scheme such as Siena, Jedi, and Rebeca [6].

The inherent limitations of wireless network makes the messaging system suitable to operate as it repeatedly tries to transmit message until it has been sent. The basic concepts in a messaging technology revolve around the key terms of message, channel and routing messages, and they will be extensively used. Transmitting data in sending messages back and forth has many advantages in a distributed application system. Some of the major advantages [7] are:

Asynchronous communication - in asynchronous communication a sender doesn't need to wait for the response to come in order to send the next request.

Throttling - a problem with messaging in Remote Procedure Calls (RPC) is that the receiver may crash due to the overhead of incoming messages. A messaging system has control on the number of requests to be sent to the receiver to process which saves the receiver from crashing.

Reliable communication - messaging system uses a store- and -forward style in providing a reliable delivery of messages.

C. Pub-sub in mobile environments

There are several papers that analyse the existing Pub/Sub model and his implementation mostly for the content-based subscription and suggests more enhanced approaches based on various optimisations. These approaches can be adapted into a mobile environment considering mobility issues of Pub/Sub system elements.

In [1] was proposed a middleware approach for a Pub/Sub implementation and its adaptation for a mobile environment. The authors explains how an event broker as a mediator can facilitate Pub/Sub communication in both centralized and decentralized mobile environments and proposes an algorithm for an optimized wireless network communication. The paper addresses the challenges of mobile networks in terms of network disconnection at any certain point and suggests the replication of users' subscription over multiple event brokers in order to improve the availability and reliability of the system in a mobile environment.

A scalable decentralized peer-based subscription approach implementation of Pub/Sub system has been proposed by authors of [8]. The study presents a topic-based deterministic information dissemination scheme that provides transparency for publisher and subscriber.

Another content-based Pub/Sub middleware approach has been proposed in [9]. The concept of mobility has been segregated into two parts – the physical mobility and the logical mobility. Depending on logical mobility, a new approach of 'location dependent subscription' using location-dependent filter has been introduced by author. In addition, the goal of [9] is to support mobile client applications in a decentralized Pub/Sub environment where clients are connected to one of the interconnected access points that serve as message routers in a distributed network. The paper implements a 'mobility support service' that provides this support to a mobile client by introducing independent mobility service proxies running at the access points of the Pub/Sub system.

A logical orientation scheme in subscription model also ensures a space optimized information Two key problems that arise in mobile applications in Pub/Sub system that have been addressed in [10] are namely scalability, in supporting large number of mobile clients and adapting to application topology as mobile components are subject to change their locations. TOPSS and JEDI are two examples of Pub/Sub systems that address scalability issue by implementing an efficient filtering mechanism at the event broker.

Although different implementations of mobile Pub/Sub systems have different prototypical and standard approaches, the common goal in all of these implementations is achieving an *efficient data dissemination strategy*. The objective of data dissemination is to transfer dynamic information (state) changes as a consequence of publishing new data and updating existing data from publishers to mobile consumers [11].

In today's heterogeneous networks that consist of Wi-Fi, 3G or 4G networks, most of the client consumers in Pub/Sub systems are smart phones and tablets, running native apps or mobile Web apps. From the developers perspective it is a controversial issue when it comes to developing apps for mobile devices. Native apps are developed solely for mobile devices which are accessible via specific device platform such as Android, Blackberry and iOS with a full access capability into the core device features. Mobile Web apps on the other hand provide the platform for single code based solution to be deployed on mobile devices with similar and more improved user experience as native apps. Thus , the mobile web app design reduces the cost of building and maintenance of mobile centric applications, and the mobile browser pattern has become the de facto standard for mobile applications since the Web is everywhere.

One key benefit of adopting mobile web methodology is the use of the latest HTML5 oriented web technology frameworks. Web frameworks such as Phone Gap and Sencha [12] support diverse mobile operating systems and allow mobile web developers to leverage their web technology skills in creating appealing applications. Moreover, these frameworks facilitate dynamic access capabilities to the device native features.

D. Web communication techniques

As a result, mobile web applications nowadays are gaining much popularity among the applications developers across several device platforms as well as in Pub/Sub system environment in disseminating information. Two of such strategies are: pull and push. In the pull approach, communication is initiated by information consumer whereas the push approach relies on information producer in initiating the communication [11]. Several web technologies are found to implement pull and push strategies. Three of such strategies expressed in their counterpart technologies, are conceptually known as *polling, long-polling and Web Sockets*. A real-time web application must receive up-to-date information. When the client browser (consumer) sends HTTP requests to the server (publisher) over a TCP connection, server acknowledges the request and issue a response back to the client.

Polling is one technique introduced in delivering real time information, in which, the client browser sends HTTP requests to the server at a regular time interval and every time the server receives a request, responds back to the client. This approach is suitable in a situation when the server update interval is known to the client so that the client can be synchronized to send request to the server based on the exact interval of message delivery. There is also a growing need for asynchronous communication in collaborative applications where multiple users interact real-time among themselves. To response to this need, the Ajax technique has been introduced which enables web browsers to fetch dynamic information from the server asynchronously using in-built JavaScript functionalities such as XMLHttpRequest. However, although Ajax solves the problem of collaborative communication, its intense communication with the server causes significant overhead especially when using the polling technique. As it is difficult to predict update interval of message dissemination in real-time application, polling data from the server with a long interval can make the communication slower whereas polling data with a short interval can result in many unnecessary HTTP requests with empty responses which causes lots of unnecessary HTTP responses.

Long-polling addresses the limitations of polling by avoiding sending request in an interval. In long-polling, as the browser initiates a HTTP connection with a server, the server maintains the connection persistently for a certain period of time and pushes the update message to the client whenever it becomes available. If the update is not available within the set

period of time, the server sends an empty response message as it times out and the connection is terminated. The browser then has to re-open another HTTP connection to send the next update request. In the asynchronous long-polling operation; the server can push update messages to the browser without the client prompting. However, performing long-polling in a groupware application where data is constantly updated will result in no improvement over the traditional polling technique as long-polling throttles the connection with lots of intermediate requests that consumes server resources [13].

Web Socket Technology. One of the latest web technology concepts introduced in the HTML5 standards as a new approach for the next generation web communication is Web Socket. It provides a *full-duplex bi-directional asynchronous* communication channel between web browser and web server applications over a single TCP socket per end point [14]. In addition, it has added the socket functionality to the browser to eliminate many problems of existing technologies. The complete Web Socket standard is the combination of the Web Socket API and the Web Socket protocol.

The Web Socket API is a draft specification standardized by W3C [14]. The API defines a communication interface between the web application and the browser [13]. The browser must expose the API to the web application so that when initiating a Web Socket connection the application invokes the following API to create a Web Socket object. Using the object, application then invokes the Web Socket API functions to open and close connection as well as send and receive messages. Current the browsers that support Web Socket standard are Firefox 6, Google Chrome 16, and Internet Explorer 10 [14]. *The Web Socket protocol* has been designed to improve the existing HTTP connection. Two primary tasks that this protocol performs are establishing connection through handshake and transferring data. The initial handshake starts with a HTTP protocol. In the browser request, the GET method indicates the end point of the connection. The Web Socket server uses values from headers sec-Web Socket-Key to calculate a hash value and send it to the client to prove that the handshake was received and sec-Web Socket-accept header field indicates whether or not the server accepts the connection.

Once the handshake between the client and the server is successfully established, the connection is ready for data transfer. In the Web Socket protocol, data is composed of sequence of frames which can be of type texts, interpreted as utf-8 text, binary data and control frame. Control frames are texts that are intended for signaling the connection for instance when the connection should be closed. Since the Web Socket protocol uses a HTTP compatible handshake, it can also use a HTTP port as well as an underlying TCP protocol for network communications. Several web-based systems are found nowadays are using the Web Socket API and the protocol as the key implementation tool. A web-based control application using Web Socket is proposed in [15] that shows how a Web Socket-based application can be built with just HTML5 without using any add-ons in the web browser. Another work by [16] integrates the Web Socket API into an existing framework to support distributed and agent-driven data mining in an enterprise environment. The work is similar to R- Web Socket except that it implements both the client and the server side interface for Web Socket API and the implementation uses Grizzy framework to provide scalability to the underlying infrastructure.

E. Application development patterns

A good architectural pattern in developing software applications can ensure a better performance for resource constraint mobile device. In talking about application design, we often encounter the term 'MVC' which is a short form of Model-View-Controller. An architectural design that is based on MVC produce a clear abstract framework in the system development process. This provides a clean separation between software components. An evolved version of MVC is MVP, stands for Model-View-Presenter that focuses on improving the presentation logic/UI logic. Unlike MVC, the *Presenter* component in MVP contains the user interface business logic of the *View*. Communication between *View* and *Presenter* thus happen through a view interface. As the UI logic of the *View* is dedicated to the *Presenter*, a direct request from *Presenter* to *View* becomes possible. *Presenter* can trigger the *View* updates without visiting though the *View* component. This is often considered as a reason in taking MVP pattern most suitable for web-based architecture. The separation of concern in presentation logic helps *Presenter* to ignore implementation details of the *View* and only concern on the method to invoke of the *View* interface. This feature of MVP provides a higher level of abstraction which made it a successor to MVC.

The traditional web application supports sequential flow of data where user had to fill a form and submit before showing the html content on the page. With the advent of AJAX, the modern UI of MVC/MVP supports *event-driven style of data flow*. As the stream of events arrives, the job of dispatcher is to determine the event type and pass it to the handler that can handle events of that type. In a client-server interaction, dispatcher and the event handlers may reside in the server side In that case, events from client's requests are queued up before transmitting them to the server to be processed. In event-driven, programs are like multiple individual modules that can be triggered based on the event types. The program is designed as a continuous loop that keeps listing for event and calls the event handler (also known as callbacks) that matches the event type.

F. Cloud computing

The foundation of cloud computing is seen as a remarkable way in consuming web services in resource poor of mobile device by offloading resource intensive computation and data storage outside the device into resource rich remote machines [18]. Computing in the cloud also provides scalable hosting of IT backend services. Several approaches have been proposed by myriads of research studies for the effectiveness of offloading techniques. Since the wireless signal may attenuate due to device mobility, these studies offer a notion of *dynamic offloading* that is said to be feasible in such network environment. [17] offers a cloud infrastructure that seamlessly offloads execution from mobile device to a replicate copy of mobile application software running in the virtual cloud server. This approach of migrating computation from a device

to a device replica gives mobile user an illusion of using powerful, feature rich device and also known as Clone Cloud. Similar approach is proposed by [16]. This study proposed to locate the cloud service software on a nearby resource-rich computer(s) called *cloudlets* that is well connected to the internet as well as to the mobile users. The approach of bringing the cloud virtual machine close to the mobile users is considered latency optimized in terms of latency and data transfer cost. In offloading mechanism, a fine grain offloading approach has shown in MAUI system [19] where instead of offloading on the whole application software, which methods to be executed remotely are decided in the runtime and thus saves energy and increase the battery life of mobile devices. Combining *cloud computing and RESTful Web services* provides a new paradigm of mobile computing. In his research specifies REST as a suitable architectural platform that lends itself well in consuming cloud Web services in resource constraint mobile device.

From the literature review, it can be concluded that the *channel based Pub/Sub* is an ideal model for a distributed system where applications are disparate and dispersed over the network. The space decoupling nature of Pub/Sub enabled mobile applications and the interacting parties who use these applications to be anonymous and independent from each other. Publisher can publish events at any time without blocking themselves and subscribers are notified asynchronously through a callback. Publisher doesn't hold any reference of subscriber which let the publisher to publish events even when the subscriber is disconnected. This decoupling in production and consumption explicitly removes dependencies among the interacting participants and increases the scalability.

The communication in Pub/Sub is asynchronous that well adapts with the distributed environment such as mobile environment. On the other hand, Web services have been a great solution in integrating distributed and disparate system applications. Due to clear semantics and uniform interface and its supportability for different message formats, REST Web Services has become the most suitable approach in consuming services in mobile environment. REST avoids the single access point in consuming services and thus increases the service scalability.

Reviewing the challenges in mobile distributed environment and the proposed solutions, this research attempts to address the following open issues;

- How can we build a RESTful Pub/Sub system in mobile environment?
- How much the system needs to comply with REST and Pub/Sub features to call it RESTful Pub/Sub?
- And because of operating in mobile environment, how can we ensure a system that is fault-tolerant and yet efficiently disseminate information?

In the rest of the paper we will try to respond to these questions in terms of a design solution, a prototype implementation and set of scenarios in order to make a realistic evaluation of our framework.

III. The Framework Design

This section looks into different REST patterns in event dissemination in accordance to the challenges mentioned in problem statement and then propose a framework that is adopted for mobile clients to consume RESTful Web Services within an event-based Pub/Sub domain. The proposed framework is designed in three main layers as shown in Figure1.

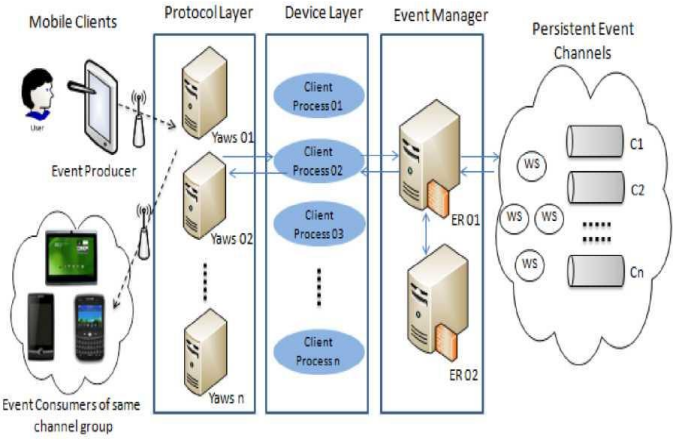


Fig. 1. The conceptual architecture

The front-end of the framework represents mobile clients who are publishers and/or subscribers of data at the Web Service (WS) channels. The backend of the framework contains Web servers as Protocol layer and Device layer, Event Manager and the cloud hosted Web Services channels. The Web servers and Event Manager act as a proxy layer between mobile clients and WS channels. Since *we adopt a Pub/Sub model*, data are disseminated in the form of events. Similarly, a mobile client that publishes events is known as the Event Producer (EP) and subscribers of these events are the Event Consumer (EC). However, an event consumer can be an event producer and vice versa. In this framework, *topic-based persistent event channels* were adopted. In topic-based persistent event channels, event producer publishes events to a specific channel topic and the event consumers show their interests for events by registering to a specific channel topic.

Event channels are collections of events represented by the event topic. In the Pub/Sub model, events are published using a single input channels that splits into multiple output channels to multicast the events to each subscriber. In the application-level, mobile client applications include User Interface (UI) layout, the business logic, and the model for managing a local storage. A stub component in the client model interacts with the skeleton of the server application. The persistent event channels are fronted with the Event Router component. , that takes the responsibility of multicasting events to the mobile subscribers. The client application includes a UI layout, the business logic and the local storage capability.

The client stub provides the functionalities of the backend server on the local device. On the contrary, the skeleton on the backend server describes the functionalities of the server application. The actual implementation of the skeleton is done

at the persistent event channel. Further, the Event Manager works as an intermediary between the skeleton and the persistent event channel. All message exchanges between the client device and the remote server takes place over the standard TCP/IP transaction layer.

TABLE I. PUB-SUB MAPPING TO REST SERVICES

Pub –Sub operation	REST model
Create Channel	POST/channel
Subscribe Channel	POST/channel/channel/topic/subscribe
Publish Events	POST/channel/channel_topic/publish
Read Events	GET/channel/channel_topic_eventMessages
Request for Updates	HEAD/channel/channel_topic
Unsubscribe Channel	DELETE/channel/channel_topic/unsubscribe

According to the Richardson's Maturity Model (RMM) [8], a RESTful dissemination of data can take four different patterns based on REST Web Service's maturity level also known as the glory of REST.

In the context of the proposed framework in this thesis, the patterns are hereby discussed as follows:

Pattern A: Using HTTP POST (Level 0) - Event-dissemination of this pattern follows level 0 of the RMM. In this pattern, services are exposed using one URI; and consumers can access the URI using a single HTTP POST method. This is similar to SOAP based WS where requests are sent to one URI and XML payloads are exchanged between the sender and receiver.

Pattern B: Using HTTP GET or POST (level 1) - Event dissemination of this pattern is based on level 1 of the RMM. In this pattern, a service is exposed as many logical resources with unique URIs contrary to single resource/service of level 0 (pattern A). A request is sent either using HTTP POST and/or HTTP GET. In this pattern, operations can be performed using HTTP POST. Sometimes HTTP GET is used in addition to HTTP POST. However, HTTP verbs do not strictly follow HTTP rules or REST constraints in this pattern.

Pattern C: Using HTTP CRUD Operations (level 2) - Services in this pattern host numerous URI-addressable resources. Unlike level 0 and 1 of the RMM, coordinating interactions in this pattern utilizes all the HTTP verbs (GET/retrieve, POST/create, PUT/update, DELETE/delete) in performing the CRUD operations. A response message in this communication utilizes the http status code

Pattern D: Using Hypermedia (level 3) - Pattern D is similar to pattern C in a way that it utilizes all the HTTP verbs in performing the CRUD operations except that it also utilizes the hypermedia element of the HTTP stack of the Web technology in the response message**.** Consuming services in a Pub/Sub framework can be challenging when complying with REST features described. This section describes how interactions can take place in REST-based manner in the proposed Pub/Sub based framework. Interactions between Web services and the service consumer are described in terms of major functionalities provided by the Pub/Sub service (as in Table 1).

A. The backend architecture

The backend server is responsible for hosting Pub/Sub Web Services. Web Services enables clients to create event channels (event groups) and publish events to the channel, subscribe to the channel(s) of their interests, be notified for resource updates of the channel and also unsubscribe from the channel. The system architecture takes a centralized topic based Pub/Sub model. The major functional components of the framework backend are shown in Figure 2, and further discussed.

Protocol and Device Layer. When an event is published in the event channel, it needs to be propagated as an update notification among respective subscribers. A published event is composed of event type, *etype*; published time, *etimestamp* and event messages, *emessage* (payload).

A published event is received by the Listener before it is transferred to the Event Manager (EM) process. It contains separate request handler for compatible transport mechanism.

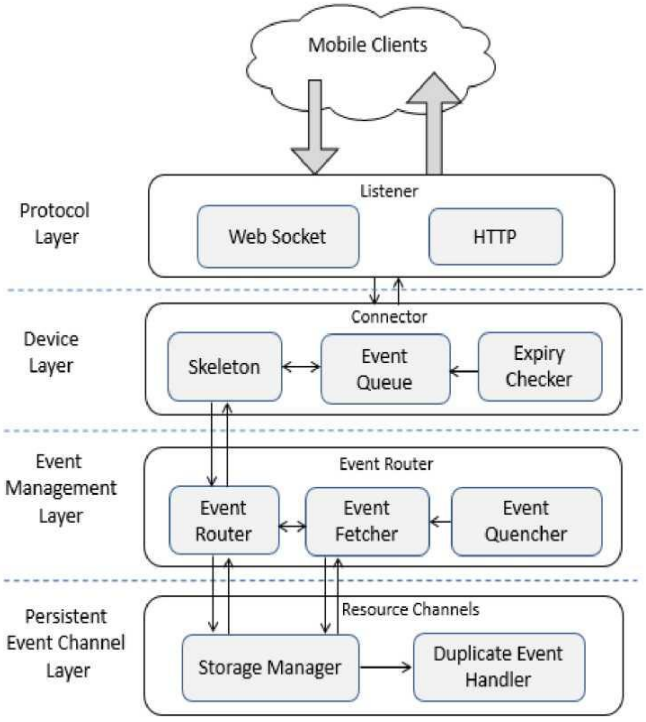


Fig. 2. Pub/Sub Backend Components

The expected transportation mechanism is the standard HTTP connection and/or Web Socket connection. Since mobile clients are using different types of device platforms, the embedded browser of native device application may not support either of this connection at any given time. To provide device transportation compatibility, a Listener process manages the request handlers for both HTTP and Web Socket. The device layer is responsible for redirecting client requests to the web services for appropriate operation execution using the connector process. This helps mobile consumers to maintain a presence at the proxy when they are disconnected

and thus resume the interaction with backend once the connection has been restored.The skeleton component of device layer provides the interface layer for Pub/Sub service, describing the functionalities that the service provides.

The Event Queue (EQ) component of the device layer buffers event update notifications received from Event Manager. It also handles duplicate event notifications to cope with network inconsistency. Event notifications are buffered in the queue until it has been propagated to the client device in FIFO style. An event is persistently removed from the queue once it is delivered to the consumer. Notifications in the Event Queue might become obsolete when event consumer is disconnected for relatively a long period of time. An event that is too old than the expected event longevity, need to be discarded from the event queue. The Expiry checker in the layer does a periodical checking in the event queue to ensure that no event notification in the queue is obsolete. Device layer also stores event data into the process storage based on their notification ID.

Event Manager (EM). The Event Manager is responsible to route event notifications to all the users who are subscribing to the channel group. Once an event is published to the persistent event channel, Event Manager invokes the Event Fetcher (EF) to fetch the list of all subscribed users of that channel. Consequently, the Event Router (ER) is invoked to actually send event notifications to the users from the subscription list. Dissemination of event updates takes a broadcast approach in delivering data to all currently active subscribers.

The Event Manager is also responsible to discard published events that arrives and does not match with the existing channel groups. An unmatched event is discarded when they are received at Event Manager. According to [Huang, Y., Molina, G., 2001], this approach is also known as event quenching. Discarding unmatched events considered to be advantageous as it does not require Event Manager or any of its replica (if any) to attempt transmitting irrelevant data to the persistent event channel over the network. Moreover, accomplishing this task at Event Manager also reduces computational workload at Event Channels.

Persistent Event Channel (PEC). The Persistent Event Channel handles consumer's request for subscribing to the channel, unsubscribing from the channel, publishing event messages to the channel and also delivering event from the channel. Event Channels maintain persistent data storage for event messages published by event producer. All published event requests are sent to duplicate event handlers to check for duplicate event messages to avoid network connection delay. This can be done by checking the event ID that has been assigned by event producer's application. An event with unique event ID is stored in the channel storage persistently. Each event in the channel is uniquely identified by its URL. And thus each event resource can be accessed my consumer by sending http requests using the standard http verbs such as HEAD (meta-data), GET (read), PUT (replace), POST (create and write).

B. *The mobile client*

In this architectural framework, mobile clients are thin clients such as smartphone and tablets. Applications for these devices are responsible to register themselves to a particular channel group or group of channels based on the channel topic by consuming the Pub/Sub web services hosted in the code. Once a device registers itself, it continues to receive event notifications for any updates made in the persistent channel. In order to provide code flexibility and interoperability, the client side application is designed following the Model-View-Presenter (MVP). A stub component of the backend server is hosted in the Model. The stub is responsible for all incoming and outgoing transactions.

Once an event update arrives at the stub, the latter passes the event to the View's logic through the Presenter to be displayed on interface layout. Likewise, event messages produced by client actions (e.g. button click) are passed to the stub through the Presenter which then transmits the data to the backend server.

The *Model* component of the client application is designed to contain a persistent storage for event notifications. Moreover, it contains a queue for unpublished events; events that are produced by the client actions but could not be delivered due to the connection loss. These unpublished events are removed from the queue once they are delivered to the backend server. All interactions between the *Presenter* and the *Model* take place though the stub. The major functionalities of a stub are as follows (see Figure 3):

- *Connection service.* The stub is responsible to connect mobile application to the proxy server. Whether the communication should take place over WebSocket connection or should it be http polling are decides by the stub.

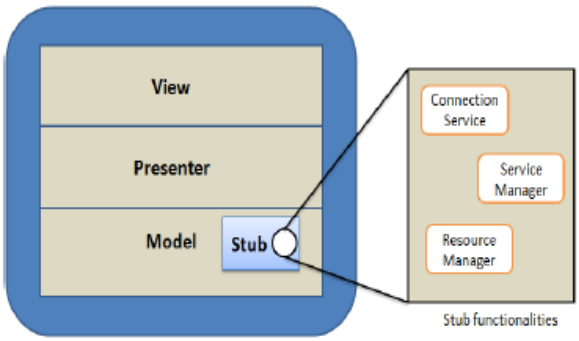


Fig. 3. Mobile client architecture

Service Manager. The stub provides the same interface of the remote cloud hosted Pub/Sub web services. It binds client's application to the remote web services over Web. It also enables client applications to invoke the consecutive functionalities of the remote web services such as subscribing to Channel, publishing data, retrieving data or unsubscribing from channel in a way as if calling to local functions. All event messages generated by these actions are encoded into JSON format before they are transmitted between client and proxy.

- *Resource Manager.* The stub is responsible to store update notifications to the local storage when it arrives from proxy. States of the stored event notifications are used to check for event updates at the proxy when client application reconnects after an intermittent connection loss. Stub also checks for the unpublished events in the queue once after every connection establishment

IV. Performance Evaluation

The experiments analysis and evaluation serve to demonstrate the framework's feasibility in various event dissemination patterns and also to identify the best performing scenario. The three major components in this experiment setup include mobile users (event producer and consumer), Pub/Sub Proxy layers (Protocol Layer, Device Layer and Event Manager), Pub/Sub Persistent Event Channels.

- Mobile client: Mobile clients are running on ASUS Transformer Prime tablet. The device specifications are Android™ 4.0 Ice Cream Sandwich OS, NVDIA® Tegra® 3 Quad-core CPU, 1 GB memory and 1.3 GHz CPU Speed.
- Pub/Sub proxies: A Windows 7 desktop machine is used to host Pub/Sub proxy layers, with 64-bit Windows 7 Professional Intel® Core ™ i5-2400 CPU, 16.0 GB Memory and3.10 GHz CPU Speed.
- Pub/Sub Persistent Event Channels: A Windows 8 desktop machine is used to host Pub/Sub event channels. The hardware specification is 64-bit Windows 8 Enterprise Intel® Core ™ i5 CPU4.0 GB 3.2 Memory, GHz CPU Speed.

A. Client app performance test

The purpose of this experiment is to observe the system's performance in request/response on different client application platforms. In this experiment, three different application platforms that have been tested are Erlang client, JavaScript Desktop browser and device embedded browser. Each of these platforms establishes WebSocket connection to its backend system.

Scenario. In this experiment, 5 kb of event messages has been published from the initial sender to the Persistent Channel and 1 kb of event messages has been pushed to mobile clients by Event Router. As the event message propagates from sender to the receiver, the Round-Trip-Time (RTT) has been observed.

Discussion. Among the three client applications, the best performance is observed on the Chrome browser running on Desktop One possible reason that the app on Android WebView performs slower than Chrome browser is because WebView is linked to the Android application layer written in Java. For every activity in WebView for example JIT (just-in-time) compilation of JavaScript, the callback function is invoked. Moreover, the integration of an external framework in the application such as Phone Gap might have added an additional execution time which in turn causes performance deterioration

B. System (protocol) overhead

This test is conducted to observe the amount of overhead the chosen dissemination approaches introduces on the system in terms of latency in consuming a resource from the Persistent Channel. The chosen approaches include client pull over HTTP Ajax and server push over WebSocket. The purpose of this test is to observe the time difference and identify which approach performs better in event dissemination.

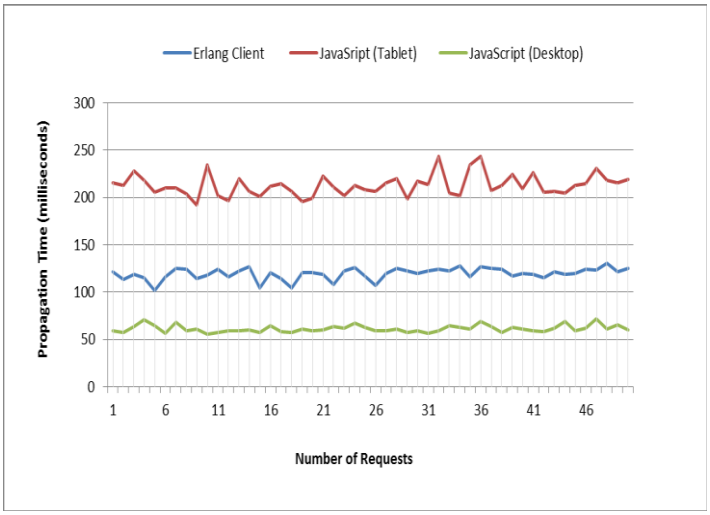


Fig. 4. RTT per request (multiple client platforms)

Scenario. The event update message is stored in the persistent channel. The experiment is conducted in two scenarios. In the first scenario, mobile consumers who are subscribing to a channel are configured to pull for event updates from the channel every 2 seconds. In the second scenario, as event updates arrives at Persistent Channel, Event Router pushes the update to the subscriber's end i.e. update propagation does not require any requests arriving from the subscribers.

Discussion. The result of client pull and server push is shown in the Figure 5. The graph shows the time for individual update propagation (50 samples) obtained from an average of five iterations where the size of each event message is 10 kb. From the graph, it can be observed that, time consumption in first scenario where the message propagates from event publisher to the server and having server send update to the subscriber as a response for update request takes much longer time comparing to the time of propagating event from publisher to the server and having server push the update to the subscriber. Time in event consumption is observed almost 1.5 times faster in server push scenario compared to client pull.

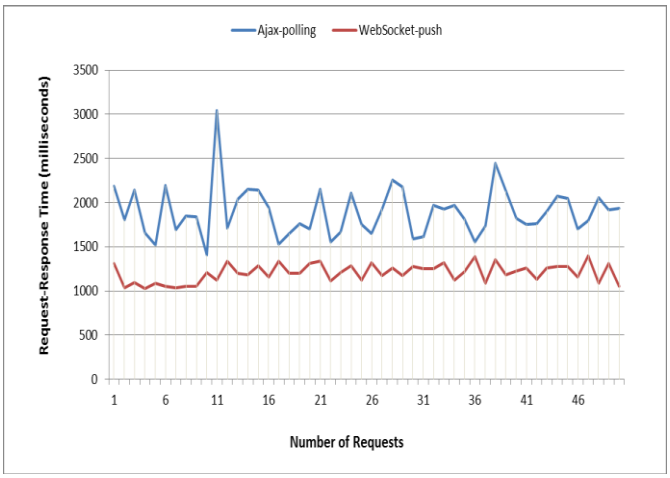


Fig. 5. Response time per request over http polling and WebSocket

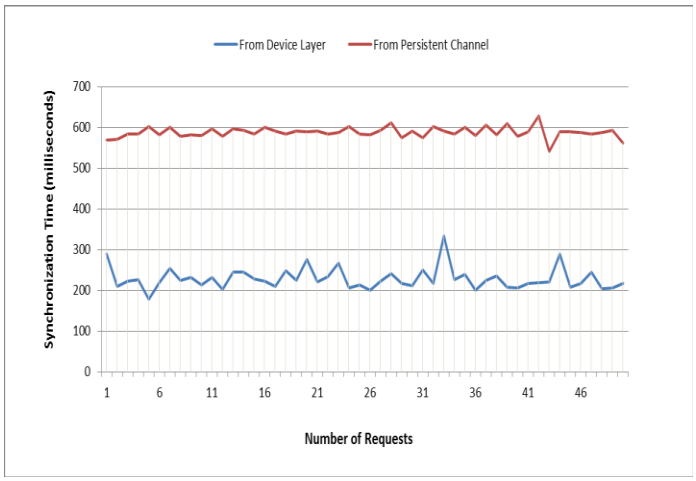


Fig. 6. Response time per request from the device layer and from Persistent Channel

C. Resource synchronization test

A framework that is designed to run part over heterogeneous network for example in this case, part over wireless network and part over LAN, one problem that arises in accessing resources from a far node is the routing overhead. In the proposed framework of this research, a client process is maintained for each individual subscriber at the device layer where the resources are stored temporarily.

If the client process is not maintained at the device layer then the alternative approach in synchronizing client side resource would be sending request for updates at the Persistent Channel which is multiple hops away from the clients. Therefore consumer's resource state can be synchronized from two different locations – Connector process of the device layer and the Persistent Event Channels. Hence, the purpose of this experiment is to observe system's performance difference in maintaining and not maintaining a client process at the device layer.

Scenario. In conducting the experiment, a resource has been published at the Persistent Channel. In first scenario, a client process with a temporary storage is maintained, hence the published resource has been pushed to the Connector by Event Router and client resource is synchronized with the backend resource at the device. In the second scenario, published resource is made available to only Persistent Channel. Hence client application is configures to synchronize its local resource at the Persistent Channel

Discussion. The results from the experiment are graphically presented in Figure 6. The graph shows the synchronization time for 50 individual requests. Each synchronization time plotted on the graph is an average time of five iterations. A resource of size 5kb has been synchronized between client's local storage and the backend storage based on client's current resource id. Results shows that the average time required to synchronize the resource from device layer is 228.5 milliseconds while it is 588 milliseconds if synchronized from the Persistent Channel Layer which is 2.6 times (157.3 %) slower. Hence, maintaining a client process in a closer proximity of the client device can result in a better performance in synchronizing data in a distributed framework.

D. Bandwith Consumption Test

This experiment analyzes the bandwidth consumption over wireless network in disseminating resource updates to the corresponding clients. The purpose of this experiment is to compare the throughput of update dissemination over traditional client pull approach with the server push based data dissemination in Pub/Sub paradigm. The experiment investigates the technique that helps in efficiently consuming available bandwidth by avoiding unnecessary network traffic in communication network. As the updates are propagated from Pub/Sub server to clients, bandwidth is calculated at server's end for every incoming and outgoing interaction.

Scenario. In this experiment, a similar scenario of System Overhead test has been adopted. This experiment is conducted in two phases. In first phase, client app is configured to send resource update request at a constant rate (i.e. every 2 seconds). Upon receiving the client request, Pub/Sub server responds with an update notification of 2kb of message payload and the updated resource. In case there is no update available, sever acknowledge the requester with a message "No update is available". In second phase, Pub/Sub server pushes the updated resource to the subscriber without subscriber prompting for the update.

Discussion. Figure 7 shows the throughput in kilobyte/second for individual resource propagation in client pull and server push approach of event dissemination. In this experiment, 10kb of data has been transferred between mobile client and server. The average throughput obtained over http polling is 5.8 kb/s when the average throughput over WebSocket is 8.6 kb/s. Bandwidth consumption over WebSocket results in at least 1.5 times higher compared to http polling.

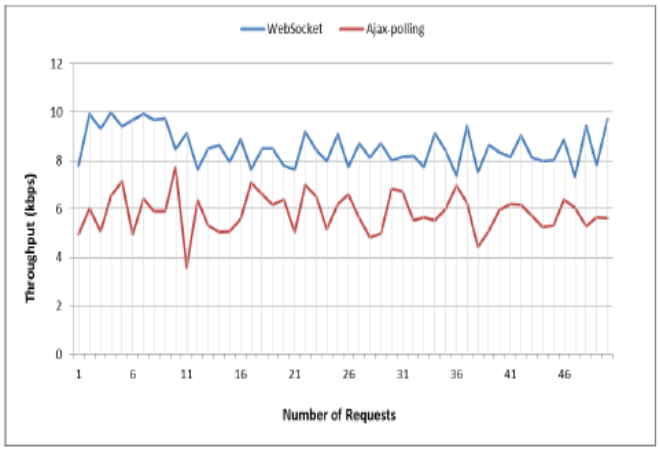


Fig. 7. Throughput per request over http polling and WebSocket

V. Conclusions and Further Developments

The research contributes in the domain of Web Services based event dissemination in Pub/Sub domain as follows: analyses different patterns of RESTful Web services within Pub/Sub domain for disseminating consumer data, studies the latest Web communication technologies and different data dissemination patterns to address the challenges of network latency in mobile environment, proposes a solution for traditional pull-based architecture by adopting WebSocket as a communication protocol and provides a platform for Pub/Sub communication on mobile environments. Further research will like to explore the following features as possible approaches that could be added to the existing framework to achieve greater performance improvements: using a decentralized Pub/Sub system, maintaining a User Profile**,** using mobile Web Service provisioning. These further developments are succinctly described in the following:

1) Decentralized Pub/Sub system. The current Pub/Sub framework is based on centralized event brokering system that relies on a single event broker. If the event broker is down then the event dissemination within the framework will be compromised hence relying on a single event broker increases the vulnerability of the entire system because it limits the system by the capacity of a single server. Hence adopting decentralized Pub/Sub model is a promising line of work. In decentralized approach, the system consists of M number of event brokers each responsible for a portion of N total subscription and hence responsible to deliver event updates to its own active subscription user's list.

*2) Maintaining a User Profile**.** The proposed framework is based on topic-based subscription scheme where users subscribe to events of a channel based on the channel topic or subject. However, subscription mechanism can be improved by introducing a subscription scheme based on the actual content of an event which provides more granularities in event subscription through offering a fine filtering mechanism on events. In this mechanism, maintaining a user profile can be useful in defining filtering rules in event subscription. Nevertheless, the proposed framework uses a flexible queuing policy where the notifications are buffered until the subscriber reconnects. A more complex and granular queuing policy would buffer undelivered notifications based on the subscriber defined propertied such as priorities and expiry dates of event channels.*

*3) Mobile Web Service Provisioning**.** One of the major trends of distributed system network and also a future direction of this research is the emergence of mobile terminals as Web Service providers also known as Mobile Hosts. When lot of research focuses on provisioning Web Services from resource constraint mobile device, some research works sees the potential of using smart and more powerful mobile devices with sufficient speed as the service delivery node in peer-to-peer settings.*

References

[1] Gartner, 2011. Gartner says Worldwide Mobile Application Store Revenue Forecast to Surpass $15 Billion in 2011.

[2] Lomotey, R.K.; Deters, R. "Reliable Consumption of Web Services in a Mobile-Cloud Ecosystem Using REST", 2013 IEEE 7th International Symposium on Service Oriented System Engineering (SOSE), On page(s): 13 - 24, vol., no., pp.13,24, 25-28 March 2013

[3] Webber, J., Parastatidis, S., Robinson, I. 2010. REST in Practice, O'Reilly Media. Retrieved on March 20th, 2012.

[4] Liu, C., Liu, Y., Ma, X., Gao, J. 2010, An Application scheme of publish/subscribe system over clustering Mobile Ad Hoc Networks. P. 1-4.

[5] Baldoni, R. and Virgillito, A. 2005. Distributed event routing in publish/subscribe communication systems: a survey. Technical Report TR-1/06. The Computer Journal, vol.50(2), pp.444 -459

[6] Fiege, L., Muhl, G. 2000. Rebeca Event-Based Electronic Commerce Architecture, Available: http://www.gkec.informatik.tu-darmstadt.de/rebeca.

[7] Hohpe, G., Woolf, B. 2004. Enterprise Integration Patterns : Designing, Building, and Deploying Messaging Solutions. Addison-Wesley, Boston. 2004.

[8] Anceaume, E., Datta, A.K., Gradinariu, M., Simon, G. 2002. Publish/subscribe scheme for mobile networks, in: Proceedings of the ACM Workshop on Principles of Mobile Computing 2002, pp. 74-81.

[9] Fiege, L.,Gartner , F.C.,Kasten , O., Zeidler , A. 2003. Supporting Mobility in Content-Based Publish/Subscribe Middleware, p.103-122.

[10] Cugola, G.,Jacobsen, H. 2002. Using publish/subscribe middleware for mobile systems.Mobile Computing and Communications Review 6(4): 25-33.

[11] Muhl, G., Ulbrich, A., Herrmann , K., Weis, T. 2004. Disseminating Information to Mobile Clients Using Publish-Subscribe. IEEE Internet Computing 8(3): 46-53.

[12] PhoneGap. 2012. Available: http://phonegap.com/

[13] Lubbers, P., Greco, F. 2010. HTML5 Web Sockets: A Quantum leap in Scalability for the Web. Available: http://soa.sys-con.com/node/1315473.

[14] WebSocket.org 2012. Available: http://www.websocket.org/

[15] Furukawa, Y. 2011.Web-based Control Application Using Websocket, ICALEPCS2011, p.673- 675.

[16] Cassetti, O., Luz, S. 2011. The WebSocket API as supporting technology for distributed and agent-driven data mining. Available: http://www.scss.tcd.ie/~casseto/NGDM11-websockets.pdf.

[17] Satyanarayanan, M., Bahl, P, Caceres, R., Davies, N. 2009. The Case for VM-Based Cloudlets in Mobile Computing, IEEE Pervasive Computing, vol. 8(4), pp. 14-23.

[18] Chun, B. G., Maniatis, P. 2009. Augmented Smartphone Applications Through Clone Cloud Execution, in Proceedings of the 12th Workshop on Hot Topics in Operating Systems (HotOS XII), May 2009.

[19] Ashik, K., Kazi, R., Deters, D., 2012, "Supporting the Personal Cloud", IEEE Asia Pacific Cloud Computing Congress 2012, Shenzhen, China, November 14-17, 2012.

A Coding Technique Based on the Frequency Evolution Creates with a Time Frequency Analysis a New Genome's Landscape

Imen MESSAOUDI, Afef ELLOUMI
Université de Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis, LR Signal, Images et Technologies de l'Information, BP 37, le Belvédère, 1002, Tunis, Tunisie.

Zied LACHIRI
Département de Génie Physique et Instrumentation, INSAT, BP 676, Centre Urbain Cedex, 1080, Tunis, Tunisie.

***Abstract*—In recent years, considerable effort has been devoted to study the biological data sets within the framework of the genomic signal processing field. However, the enormous amount of data deposited into public databases makes the search for useful information a difficult task. Effectively, the choice for a convenient analysis approach is not at all obvious at all. In this work, we provide a new way to map the genomes within the form of images. The mapping uses the Complex Morlet wavelet as analysis technique and the Frequency Chaos Game Signal (FCGS) as digital dataset. Before processing the wavelet analysis, we build the FCGS in such a way that we can follow the frequency evolution of nucleotides' occurrence along the genome. The time-frequency analysis of the FCGS signals constitutes a pertinent tool for exploring the DNA structures in the *C.elegans* genome-wide landscape.**

Keywords*—*C.elegans; Complex Morlet wavelet; Frequency Chaos Game Signal; Genome exploration

I. Introduction

With the advances in the field of genomics, the sequencing techniques keep improving; which speeds up the collection of the biological data. Consequently, the amounts as well as the types of the data are continually increasing. Hence the need for new tool that permits an easy navigation within the genomes. Nowadays, researchers in genomics rely on a standard graphical representation of chromosomes called ideogram. Ideograms allow genomic data visualization using points, lines and other shapes to indicate the location of particular sites along the chromosomes [1]. However, the ideograms' annotations must be updated once one discovers new DNA hotspots within the chromosomic sequences. Thus, it is better to find other tool that permanently describes all the chromosome structures independently from their complexity. The idea consists in finding an adequate representation tool that directly maps the DNA produced by sequencing (i.e. in its character form) [2] [3] without a need for biological experiments or alignment algorithms [4] [5] [6] or automatic pipelines [7] to annotate all the inherent components.

II. Related Work

Within the context of the genomic signal processing, the joint time-frequency analysis based on the Fourier transform has played a key role in the data characterization and visualization [8-11]. In fact, the color spectrograms were shown to provide significant information about periodicities and recurrent motifs along the bio-molecular sequences [8] [9]. Furthermore, the tri-color spectrograms, which are obtained by the reduction of dimensionality, give a unique visual signature of specific regions of the DNA [12]. Nevertheless, the problem with STFT goes back to its limited time-frequency resolution capability. Indeed, according to the Heisenberg uncertainty principle, one cannot obtain a good resolution in both the time and the frequency domains due to the fixed STFT window's length. Thus, the wavelet transform appears to be a good solution to overcome the STFT resolution limitation [13-17]. Given the localization wavelet properties, recent works were oriented towards investigating the DNA correlations [18-20] as well as the identification of the coding regions in genes [21] and some of the repeating protein motifs [22]. In this framework, we propose the continuous wavelet analysis to map the genomic DNA as scalogram images based on the complex Morlet wavelet. The main objective of this work consists in unraveling the localized spectral behaviors of different DNA structures. Since the DNA sequences are stored in the biological databases within the form of strings, it is necessary to convert them into numerical data; which will enable in turn the signal processing based applications. This operation defines the so-called "DNA coding". Here, we are concerned with a new numerical assignment scheme, which is the Frequency Chaos Game Signal (FCGS). The latter, allows following the evolution of the oligomers frequency of occurrence. Furthermore, by combining the FCGS with the wavelet analysis we create a new perspective to represent the information within any given genomic sequence. The method is not only effective and original; it also enhances specific information about the sequence at each FCGS level. To prove the efficacy of the work in terms of the genomes landscapes visualizing, we consider the *C.elegans* genome.

III. Mapping The DNA Sequences: From Character Strings Towards The Frequency Chaos Game Signal

The Frequency Chaos Game Signal is a new DNA coding technique consisting in a linear form of the Frequency Chaos Game Representation (FCGR). The latter method illustrates a DNA sequence in the form of a 2D image [23-25], based on the Chaos Game theory [26] [27]. The method (we mean the FCGR) encodes the oligomers frequency of apparition according to given color scale; where each frequency occupies

a precise emplacement in the representation area. Consider the following sequence:

Seq_{comp}=ACGATACAGATCAGATTTAGACAGACCGATAGTAGACGATCAGATCACCAGTGAC.

The monomers, dimmers and timers frequencies of apparition as well as the related words organization are given by TABLE I and TABLE II.

In our coding approach, we take the frequency matrices as assignment base [28] [29]. For this purpose, we fix the representation order *k* and we generate the $FCGR_k$ for the totality of the entry sequence. For this coding we generally take the chromosomic sequence as entry data set, because we want that the FCGS reflects the statistical properties of the genome itself. Then, we read the bases' succession by a group of *k*-nucleotides and we assign to each position the correspondent frequency value. For example **seq=ACGATACA**, which is a portion of the **seq_{comp}**. So, based on the FCGRs matrices previously extracted, we attribute to the monomers, dimmers and timers the frequency values as illustrated in TABLE III.

If we want to encode the totality of the sequence **Seq_{comp}**, we do the same thing. In this case, the resulting $FCGS_1$, $FCGS_2$ and $FCGS_3$ plots are given in Figure 1.

TABLE I. FCGR MATRICES ORDER 1 AND 2 WITH THE ASSOCIATED MATRICES OF WORDS: EACH WORD HAS THE FREQUENCY OF APPARITION WHICH APPEARS AT THE SAME EMPLACEMENT

FCGR2	**$FCGR_1$**
Frequency matrix 0.0556 0.0741 0.0185 0.0185 0.1296 0.0185 0.2037 0.0556 0.1481 0.1667 0.0741 0.0370 0.0185 0.1296 0.0926 0.0556	**Frequency matrix** 0.2182 0.2182 0.3818 0.1818
Dimers organization CC CG GC GG CA CT GA GT AC AG TC TG AA AT TA TT	**Monomers organization** C G A T

TABLE II. $FCGR_3$ MATRIX WITH THE ASSOCIATED MATRIX OF WORDS: EACH WORD HAS THE FREQUENCY OF APPARITION WHICH APPEARS AT THE SAME EMPLACEMENT

$FCGR_3$
Frequency matrix 0.0189 0.0377 0.0189 0.0189 0.0189 0.0189 0.0189 0.0189 0.0377 0.0189 0.0755 0.0189 0.0189 0.0189 0.0189 0.0189 0.0377 0.1132 0.0189 0.0189 0.0943 0.0189 0.0189 0.0377 0.0189 0.0189 0.0189 0.0189 0.0189 0.1321 0.0377 0.0189 0.0566 0.0566 0.0189 0.0189 0.0189 0.0189 0.0189 0.0189 0.0566 0.0189 0.1321 0.0566 0.0755 0.0189 0.0377 0.0189 0.0189 0.0189 0.0755 0.0189 0.0377 0.0755 0.0189 0.0189 0.0189 0.0189 0.0566 0.0377 0.0189 0.0189 0.0377 0.0377
Trimers organization CCC CCG CGC CGG GCC GCG GGC GGG CCA CCT CGA CGT GCA GCT GGA GGT CAC CAG CTC CTG GAC GAG GTC GTG CAA CAT CTA CTT GAA GAT GTA GTT ACC ACG AGC AGG TCC TCG TGC TGG ACA ACT AGA AGT TCA TCT TGA TGT AAC AAG ATC ATG TAC TAG TTC TTG AAA AAT ATA ATT TAA TAT TTA TTT

TABLE III. LINEARIZATION OF THE OLIGOMERS FREQUENCIES TO OBTAIN FCGSK WITH K={1, 2 AND 3}

Monomers (seq)=	A, C, G, A, T, A, C, A.
$FCGS_1$	**0.3818, 0.2182, 0.2182, 0.3818, 0.1818, 0.3818, 0.2182, 0.3818.**
Dimers (seq)=	AC, CG, GA, AT, TA, AC, CA.
$FCGS_2$	**0.1481, 0.0741, 0.2037, 0.1296, 0.0926, 0.1481, 0.1296.**
Trimers (seq)=	ACG, CGA, GAT, ATA, TAC, ACA.
$FCGS_3$=	**0.0566, 0.0755, 0.1321, 0.0566, 0.0377, 0.0566.**

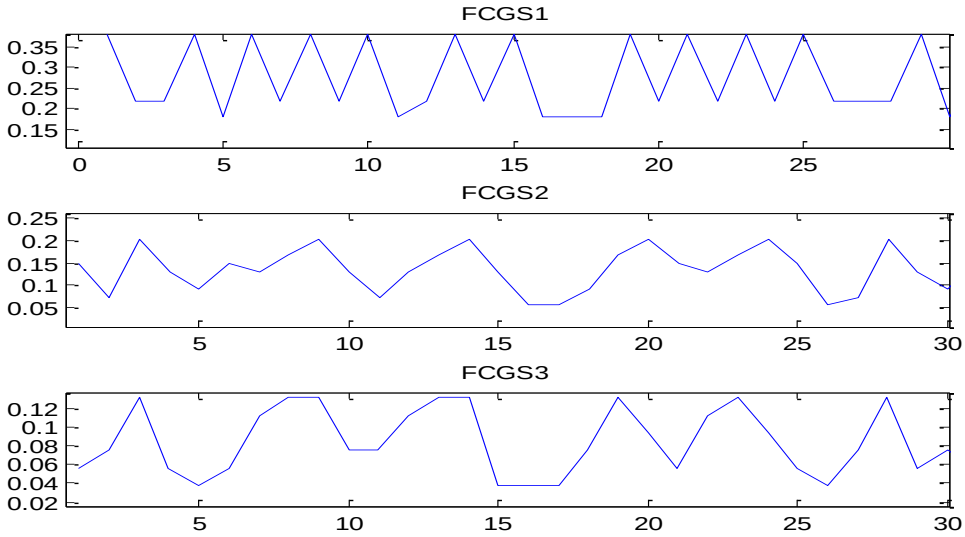


Fig. 1. Representation of the sequence "ACGATACAGATCAGATTTAGACAGACCGATAGTAGACGATCAGATCACCAGTGAC" by $FCGS_1$, $FCGS_2$ and $FCGS_3$

IV. COMPLEX MORLET WAVELET ANALYSIS

Unlike the Fourier transform, which is based on the average of signals contents within a fixed window, the wavelet transform offers very good time-frequency localization as it satisfies the uncertainty principle [30]. Indeed, the wavelet transform adapts its window (called mother wavelet and denoted ψ) in such a way it shortens at high frequencies and expands at low frequencies depending on a scale parameter a. At each scale, the daughter wavelet shifts in time using a shift parameter b to permit the convolution of the input signal with the analysis window. This principle is the basis of the Continuous Wavelet Transform (CWT). Thus, the daughter wavelets are generated by equation (1); and the continuous wavelet transform is defined by equation (2).

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right),\ a \succ 0,\ b \in \square \tag{1}$$

$$T_{\psi}(X)(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} X(t)\psi^{*}\left(\frac{t-b}{a}\right)\,dt \tag{2}$$

Here * represents the operation of complex conjugate and $X(t)$ is the input signal function. The most usual way to display wavelet transforms is to look at the absolute wavelet coefficients: $|T_{\psi}(a,b)|$, giving a time-scale representation. This representation is named scalogram. The scale set is proportional to the frequency one, it follows that one can project the wavelet coefficients in the time-frequency plan [31-33]. In general, it is preferable to choose a continuously differentiable mother wavelet with compact support [34]. In our analysis, we choose the Complex Morlet wavelet which is a Gaussian-windowed complex sinusoid :

$$\psi(t) = \pi^{-\frac{1}{4}}\left(e^{i\omega_0 t} - e^{-\frac{1}{2}\omega_0^2}\right)e^{-\frac{1}{2}t^2} \tag{3}$$

In practice, one often takes ω_0>=5 to satisfy the admissibility condition which is required by CWT [34].

V. GENOME-WIDE VISUALIZATION OF C.ELEGANS

This paper concerns the DNA representation into time-frequency images based on the complex Morlet wavelet analysis. For this purpose, we turn our attention to the *C.elegans* genome, whose sequences and annotations are extracted from the NCBI database [35]. Concerning the coding, we consider the FCGS technique with order {1, 2 and 3}. As for the wavelet analysis, we take a mother wavelet with ω_0 >5.4285. We fix, then, the number of scales to 64. Afterward, we perform the CWT on the FCGS signals for the six chromosomes. Since the scalograms yield almost the same global behavior, we selected results relating to the chromosome1 for illustration. Observing the important length of the chromosome, we proceed by zooming into the scalogram. A myriad of structures are shown to possess specific time frequency behavior which exhibits a number of

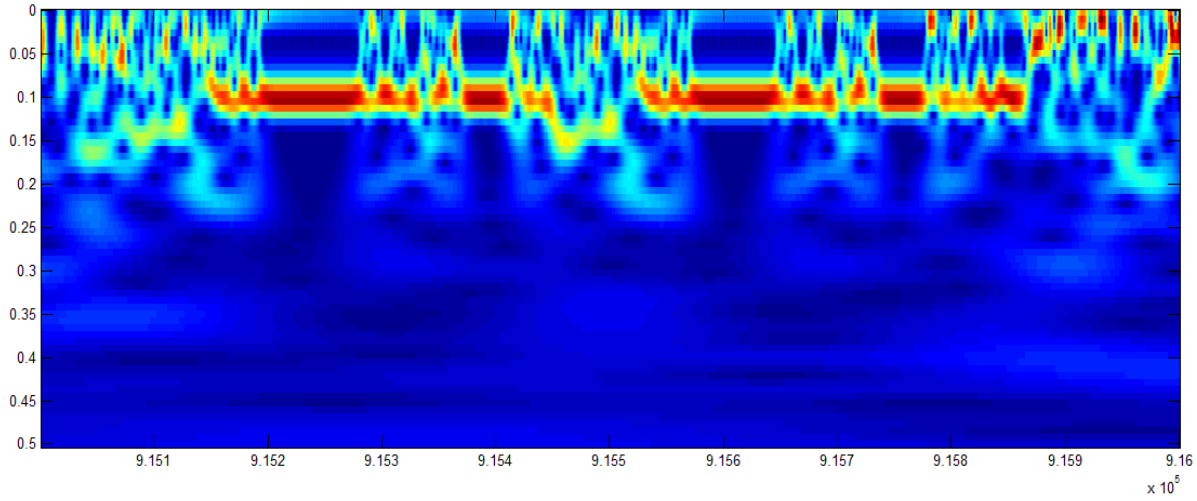
repetitive motifs. Here we are interested with (GCCTAA)n, CEMUDR1, IR3_CE, MSAT1_CE, HelitronY4_CE, HelitronY1A_CE and Helitron2_CE (see description in TABLE IV).

- The first sequence consists of two (TTAGGC)n motifs which are spread over 10^3 bp. The Figure 2 illustrates the correspondent representation while coding with $FCGS_1$, $FCGS_2$ and $FCGS_3$.

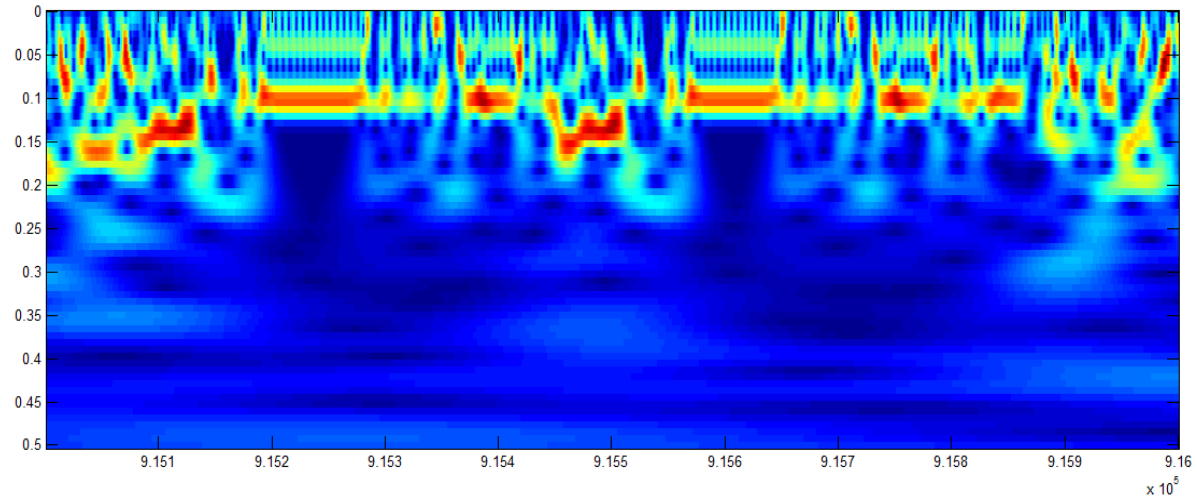
From subfigures (a) and (b), we mainly note the characterization of the structure by concentrated energy around the frequency 0.1 (which is equivalent to the 10bp-periodicity). The latter frequency is generally related to the nucleosome formation [10] [11] [36-39]. However, in the subfigure (c), energy level decreases. By comparing all the subfigures, we conclude that energy indicating the 10 periodicity decreases instead of the enhancement of other shapes within other frequency bands when we raise the FCGS order.

TABLE IV. BOUNDARIES AND LENGTHS OF THE SEQUENCES HAVING SPECIFIC TIME-FREQUENCY SIGNATURES

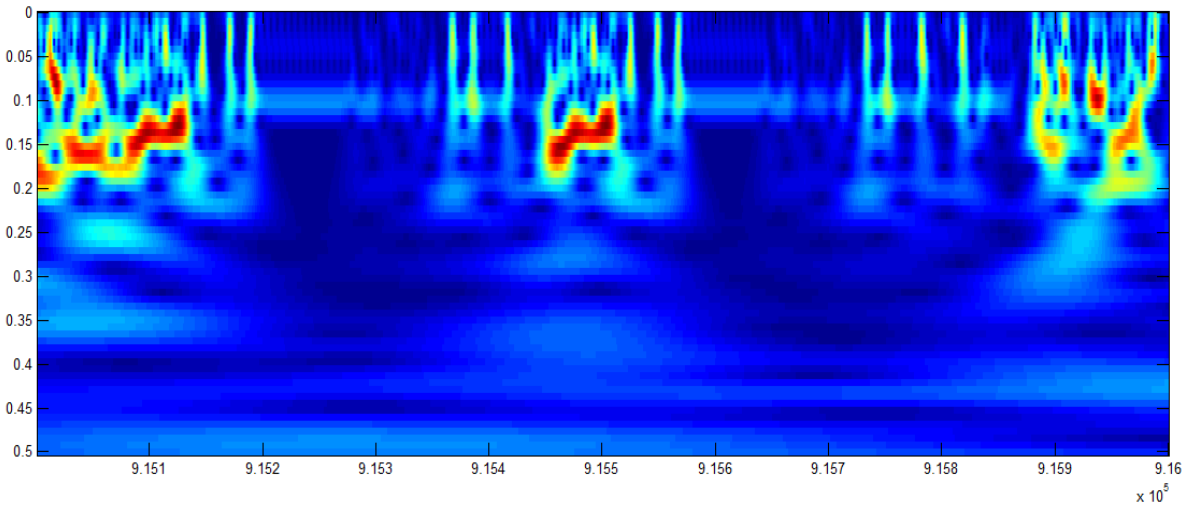
	Start	End	Illustrated Motifs	Sequence Length (bp)
Sequence n°1	915001	916000	(GCCTAA)n	1000
Sequence n°2	13935001	13937000	CEMUDR1	2000
Sequence n°3	1922001	1926000	IR3_CE	4000
Sequence n°4	9816001	9824000	MSAT1_CE HelitronY4_CE HelitronY1A_CE Helitron2_CE	8000



-a-



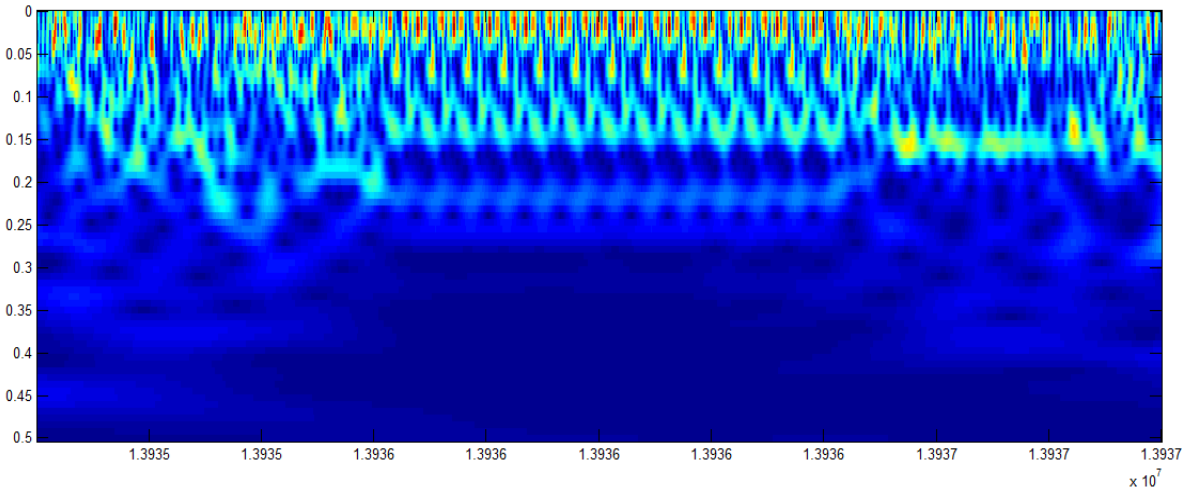
-b-



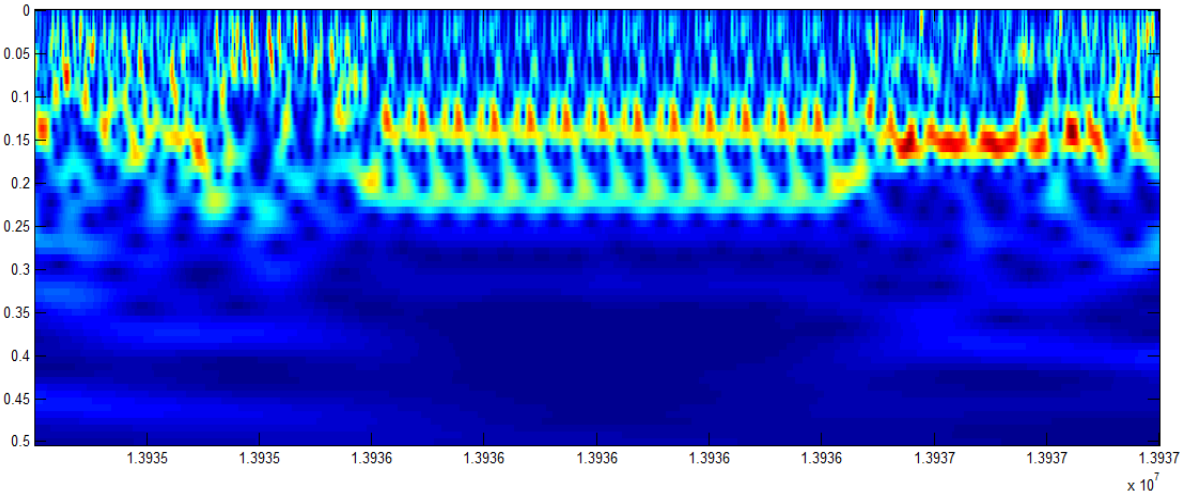
-c-

Fig. 2. Scalogram representation of two successive (TTAGGC)n sequences when coded with $FCGS_1$ (a), $FCGS_2$ (b) and $FCGS_3$ (c)

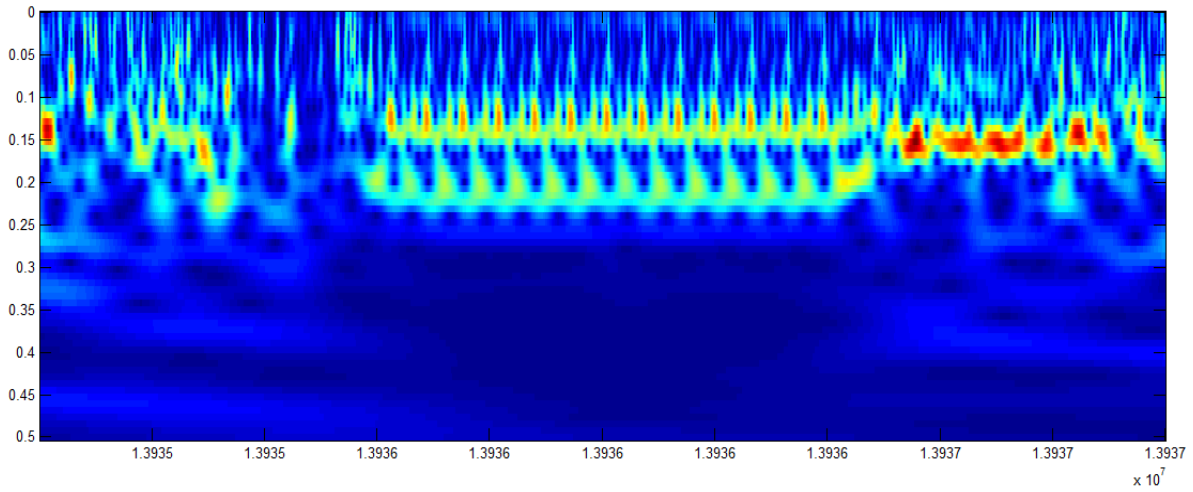
- The Figure 3 reveals the time-frequency behavior of a CEMUDR1 example. At first glance, we see periodic motifs around the frequency 0.15 (corresponding to the 6.5 bp periodicity which is generally found in non-coding DNA of genes [21]); which offers an easy way to detect the presence of the structure. A careful inspection of subfigures (from a to c) reveals that each level of the FCGS coding highlights specific repetitive shapes within a given frequency band while keeping the overall aspect.



-a-



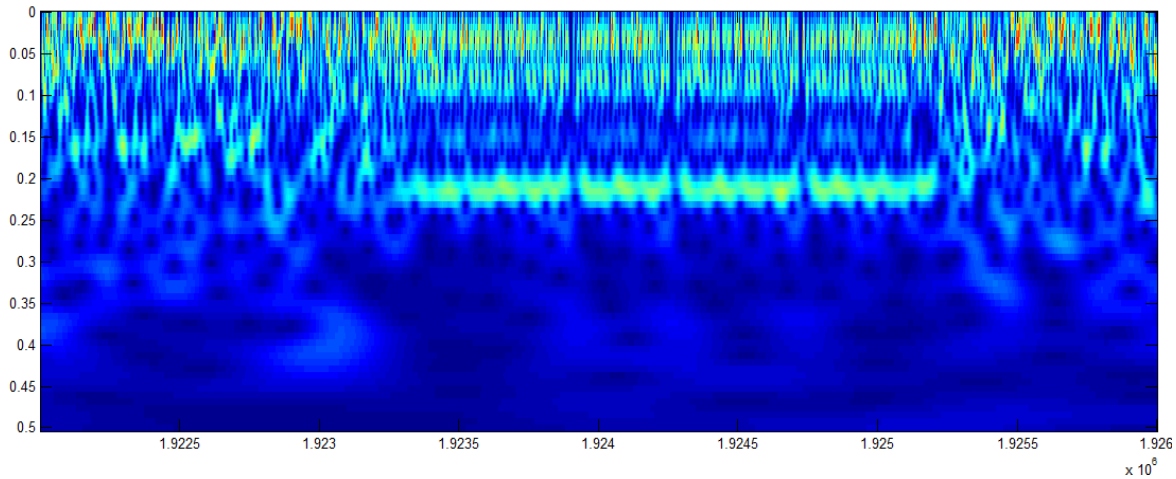
-b-



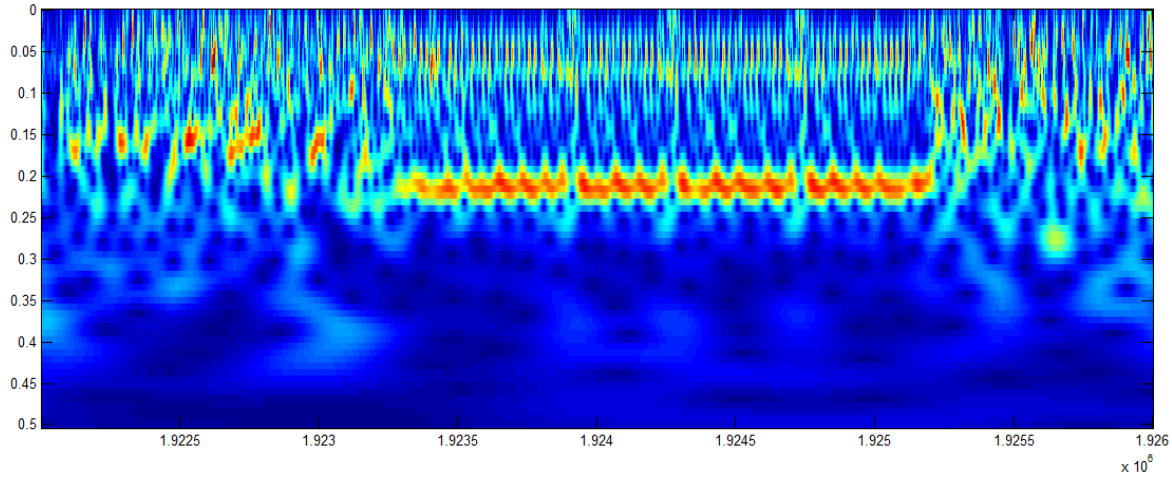
-c-

Fig. 3. Scalogram representation of the CEMUDR1 structure when coded with $FCGS_1$ (a), $FCGS_2$ (b) and $FCGS_3$ (c)

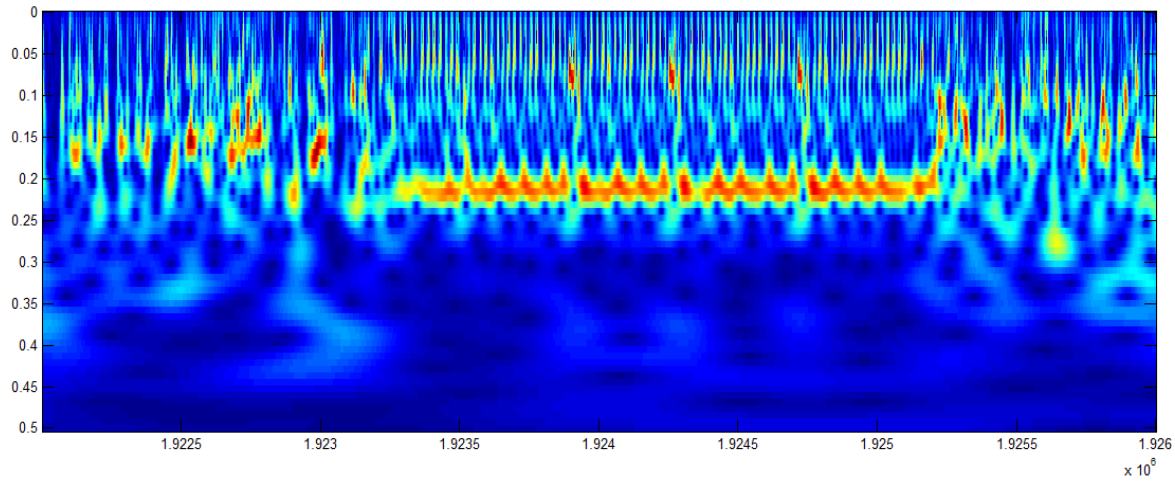
- Concerning the IR3_CE structure (Figure 4), the FCGS scalograms show characteristic signature having as main attribute the high energy around the frequency 0.2 (equivalent to the 5bp-periodicity). Moreover, changing the FCGS scale is useful for visualizing the different frequency bands that characterize the sequence for the considered order.



-a-



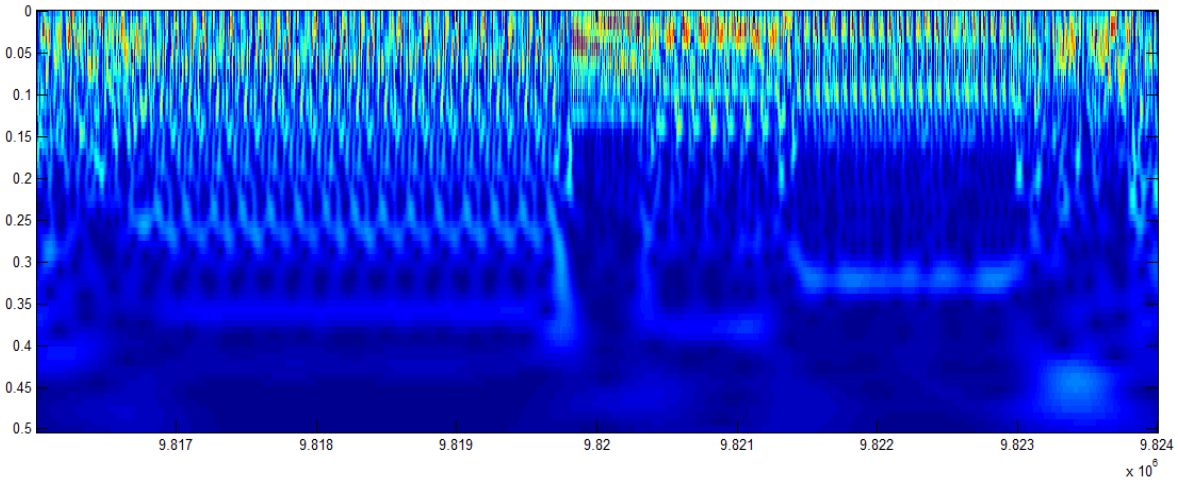
-b-



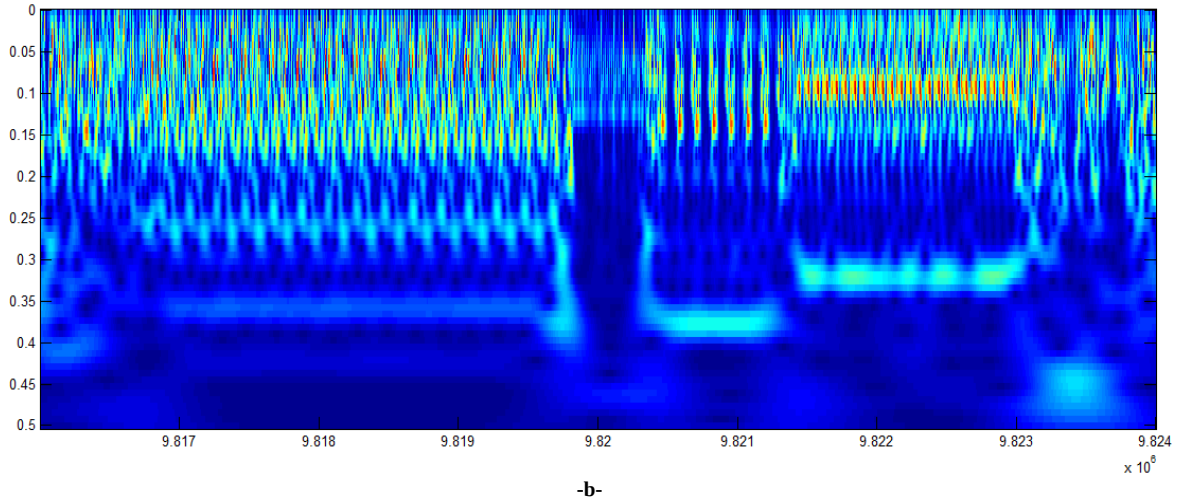
-c-

Fig. 4. Scalogram representation of the IR3_CE structure when coded with $FCGS_1$ (a), $FCGS_2$ (b) and $FCGS_3$ (c)

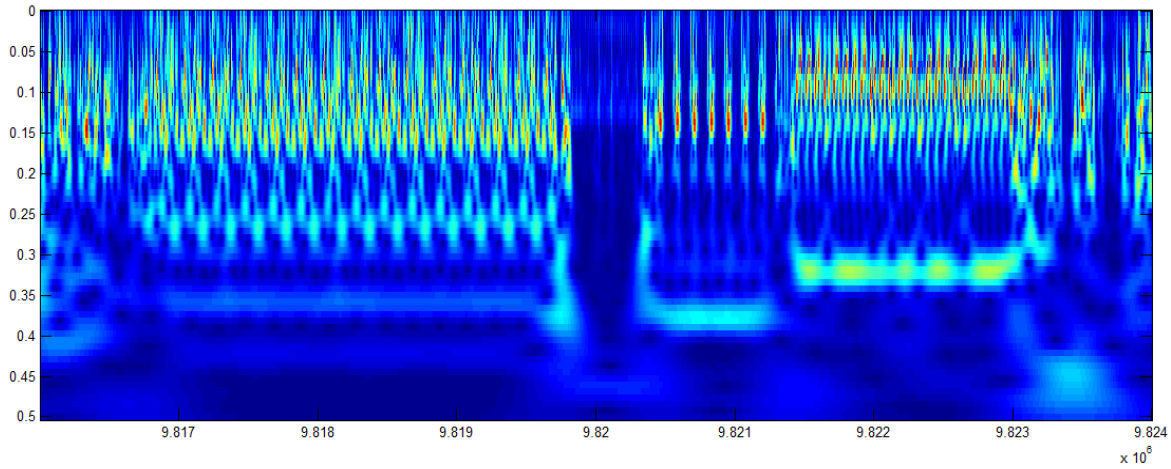
- The final example consists on a succession of a minisatellite (MSAT1_CE) and three helitrons from different classes (HelitronY4_CE, HelitronY1A_CE and Helitron2_CE). Each of these structures is shown to possess unique signature where the global behavior does not change too much when the FCGS order changes (Figure 5). However at each scale, the difference lays in the energy band repartition. Further, thanks to our representation we can easily detect the boundaries of each sequence.



-a-



-b-



-c-

Fig. 5. Scalogram representation of the succession of MSAT1_CE, HelitronY4_CE, HelitronY1A_CE and Helitron2_CE when coded with $FCGS_1$ (a), $FCGS_2$ (b) and $FCGS_3$ (c)

Apart from these examples, we succeeded in identifying other structures due to their unique time-frequency features; such as: Introns, CeRep, Satellites, Minisatellites, Microsatellites, Cele14b, Tir9ta, Longpal, Npalta, etc.

VI. Conclusion

The work presented here, described a new way to represent the genomic DNA in the form of images. Our goal was to offer visual navigation of the genome where different structures can be easily distinguished through a specific signature in the representation's plan. This has the advantage of avoiding the inaccurate and unavailable annotations as well as the long and expensive experiments. This is why; we based our study on the Frequency Chaos Game Signal (FCGS) and the Complex Morlet Wavelet analysis. The FCGS approach is a new coding tool to convert the DNA strings into signals that reflect the statistical properties of genomes. These signals are converted in turn into scalogram images that reflect the periodic features of DNA using the complex Morlet wavelet. The analysis reveals the characterization of different structures by particular periodic motifs localized within particular frequency bands with different energy levels. Furthermore, the variation of the FCGS order offers variability in terms of information while keeping the global aspect of the sequence's behavior.

VII. Future Work

Overall, our analysis is a promising and fruitful direction in the sense that it forms a strong base for classifying the DNA structures as well as recognizing unknown sequences and rectifying the available annotations. Then, future work will focus on extracting pertinent parameters for DNA structures' classification based on the generated scalograms.

References

[1] C. O'Connor, "Chromosome mapping: Idiograms", *Nature Education*, vol. 1, no. 1: 107, 2008.

[2] H.T chang , C.J Kuo, N.W . Lo and W.Z Lv, "DNA sequence Representation and comparison Based on Quaternion Number System", *International Journal of Advanced Computer Science & Applications* (*IJACSA*), vol. 3, no 11, 2012.

[3] K.S. Sathish and N. Duraipandian, "An effectice identifcation of Species from DNA Sequence: A Classification Technique by integrating DM and ANN", *International Journal of Advanced Computer Science & Applications* (*IJACSA*), vol. 3, no 8, 2012.

[4] P. Stothard and D. Wishart, "Automated bacterial genome analysis and annotation", *Elsevier, Current Opinion in Microbiology*, vol. 9, pp. 505-510, 2006.

[5] X. Huang, M. D. Adams, H. Zhou and A. R. Kerlavage, "A tool for analyzing and annotating genomic sequences", *Genomics*, vol. 46, pp. 37-45, 1997.

[6] R. Gupta, P. Agarwal and A. K. Soni, "Genetic Algorithm Based Approach for Obtaining Alignment of Multiple Sequences", *International Journal of Advanced Computer Science & Applications* (*IJACSA*) vol. 3, no 12, 2012.

[7] E. J. Richardson and M. Watson, "The automatic annotation of bacterial genomes", *Briefings in Bioinformatics*, pp. 1-12, 2012.

[8] D. Anastassiou, "Genomic signal processing", *IEEE Signal Processing Magazine*, 2001.

[9] L. Wang and L.D. Stein, "Localizing triplet periodicity in DNA and cDNA sequences", *BMC Bioinform*atics, 2010.

[10] A. E. Oueslati, Z. Lachiri, and N. Ellouze, "3D Spectrum Analysis of DNA Sequence: Application to Caenorhabditis elegans Genome", *Bioinformatics and Bioengineering BIBE, Proceedings of the 7th IEEE International Conference on*, pp. 864- 871, 2007.

[11] A.E. Oueslati, I. Messaoudi, Z. Lachiri, and N. Ellouze, "Spectral Analysis of global behaviour of C. Elegans Chromosomes", *Fourier Transform Applications*, Intech, 2011.

[12] D. Sussillo, "Spectrogram Analysis of Genomes", *EURASIP Journal on Applied Signal Pr*ocessing 1,pp. 29-42, 2004.

[13] K. B. Alsberg, , A. M. Woodward and D. B. Kell, "An introduction to wavelet transforms for chemometricians: A time-frequency approach", *Chemometrics and Intelligent Laboratory Systems*, vol. 37, no.2, pp.215-239, 1997.

[14] P.J. Mena-Chalco, H. Carrer, Y. Zana, and R.M. Cesar, "Identification of Protein Coding Regions Using the Modified Gabor-Wavelet Transform", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 5(2), 2008.

[15] J.A. Tenreiro Machado, A.C. Costa and M. D. Quelhas, "Wavelet analysis of human DNA", Genomics, vol. 98, pp. 155–163, 2011.

[16] B. Bhosale, B. S. Ahmed and A. Biswas, "Wavelet Based Analysis in Bio-informatics", Life Science Journal, vol. 10, no. 2, 2013.

[17] L. Ravichandran, A. Papandreou-Suppappola, A. Spanias, Z. Lacroix, and Ch. Legendre, "Waveform Mapping and Time-Frequency Processing of DNA and Protein Sequences", IEEE TRANSACTIONS ON SIGNAL PROCESSING, vol. 59, no. 9, pp. 4210-4224, 2011.

[18] A. Arneado, Y. D'Aubenton-Carafa, B. Audit, E. Bacry, J. F. Muzy and C. Thermes, "What can we learn with wavelets about DNA sequences?", *Physica A* 249, pp. 439-448, 1998.

[19] G. Dodin, P. Vandergheynst, P. Levoir, C. Cordier and L. Marcourt, "Fourier and Wavelet Transform Analysis, a Tool for Visualizing

Regular Patterns in DNA Sequences", J. Theor. Biol. 206, pp.323-326, 2000.

[20] A. A. Tsonis, P. Kumar, J. B. . Elsner, P. A. Tsonis, "Wavelet Analysis of DNA sequences". *Physical Review E* 53, pp. 1828-1834, 1996.

[21] P. J. Mena-Chalco, H. Carrer, Y. Zana, and R. M. Cesar, "Identification of Protein Coding Regions Using the Modified Gabor-Wavelet Transform", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5(2), 2008.

[22] B. K. Murray, D. Gorse and J. M. Thornton, "Wavelet Transforms for the Characterization and Detection of Repeating Motifs", *J. Mol. Biol*. 316, pp. 341-363, 2002.

[23] J. S. Almeida, J. A. Carrico, A. Maretzek, P. A. Noble, M. Fletcher M, "Analysis of genomic sequences by Chaos Game Representation", *Bioinformatics*, vol. 17(5), pp.429-37, 2001.

[24] A. Fiser, G. E. Tusnady and I. Simon, "Chaos game representation of protein structures", *J.Mol Graphics*, vol. 12, pp. 295-304, 1994.

[25] Y. W. Wang, K. Hill, S. Singh, L. Kari, "The spectrum of genomic signatures: from dinucleotides to chaos game representation", *Gene*, vol. 346, pp. 73-185, 2005.

[26] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, B. Fertil, "Genomic signature: characterization and classification of species assessed by chaos game representation of sequences", *Mol.Biol.E*, vol. 16(10), pp.1391-1399, 1999.

[27] P. Deschavanne, A. Giron, J. Vilain, CH. Dufraigneand and B. Fertil "Genomic Signature Is Preserved in Short DNA Fragment", *International Symposium on Bio-Informatics and Biomedical Engineering, IEEE*, pp. 161-167, 2000.

[28] I. Messaoudi, A. E. Oueslati and Z. Lachiri, "Complex Morlet Wavelet Analysis of the DNA Frequency Chaos Game Signal and Revealing Specific Motifs of Introns in C.elegans", *International Conference on Control, Engineering & Information Technology* (*CEIT*'13), 2013.

[29] I. Messaoudi, A. E. Oueslati and Z. Lachiri, "Revealing Helitron signatures in Cænorhabditis elegans by the Complex Morlet Analysis based on the Frequency Chaos Game Signals", 2nd *International Work on Bioinformatics and Biomedical Engineering* (*IWBBIO*), 2014.

[30] A .Grossmann and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape," *SIAM: Journal on Mathematical Analysis*", vol. 15, pp.723-736, 1984.

[31] M. Steinbuch and M.J.G. van de Molengraft, Eindhoven University of Technology, Control Systems Technology Group Eindhoven, "Wavelet Theory and Applications", a literature study, R.J.E. Merry, DCT, 2005.

[32] N. C. F. Tse and L. L. Lai, "Wavelet-Based Algorithm for Signal Analysis", *EURASIP Journal on Advances in Signal Processing*, 2007.

[33] A.H. Najmi and J. Sadowsky, "The Continuous Wavelet Transform and Variable Resolution Time–Frequency Analysis", *Johns Hopkins Apl Technical Digest*, vol. 18, No.1, 1997.

[34] X. P. Zhang, M. D. Desai and Y. N. Peng, "Orthogonal Complex Filter Banks and Wavelets: Some Properties and Design", *IEEE Transactions on signal processing*, vol. 47, n°. 4, 1999.

[35] http://www.ncbi.nlm.nih.gov/Genbank/.

[36] E. Segal, Y. Fondufe-Mittendorf, L. Chen, L. , A. Thamstrom, Y. Field et al., "a genomic code for nucleosome positioning", *Nature* 442, pp.772-778, 2006.

[37] J. Widom and P. T. Lowary, "Nucleosome packaging and nuclesosome positioning of genomic dna", *In Proceedings of the National Academy of sciences of the United States of America* 94, pp. 1183-1188, 1997.

[38] R. D. Kornberg, "Chromatin structure: a repeating unit of histones and DNA", *Science* 184, pp. 868-871, 1974.

[39] E. N. Trifonov and J. L. Sussman, " The Pitch of chromatin DNA is Reflected in its Nucleotide Sequence", *Proceedings of the National Academy of Sciences of the United States of America*, 77, 7, part 2: Biological Sciences, pp. 3816-3820, 1980.

On the Parallel Design and Analysis for 3-D ADI Telegraph Problem with MPI

Simon Uzezi Ewedafe
Department of Computing
The University of the West Indies,
Mona Kingston 7, Jamaica

Rio Hirowati Shariffudin
Institute of Mathematical Sciences
Universiti Malaya
Kuala Lumpur, Nigeria

***Abstract*—In this paper we describe the 3-D Telegraph Equation (3-DTEL) with the use of Alternating Direction Implicit (ADI) method on Geranium Cadcam Cluster (GCC) with Message Passing Interface (MPI) parallel software. The algorithm is presented by the use of Single Program Multiple Data (SPMD) technique. The implementation is discussed by means of Parallel Design and Analysis with the use of Domain Decomposition (DD) strategy. The 3-DTEL with ADI scheme is implemented on the GCC cluster, with an objective to evaluate the overhead it introduces, with ability to exploit the inherent parallelism of the computation. Results of the parallel experiments are presented. The Speedup and Efficiency from the experiments on different block sizes agree with the theoretical analysis.**

Keywords—3-DTEL; ADI; MPI; SPMD; DD; Parallel Design

I. INTRODUCTION

Parallel computing has greatly motivated the research works on the parallel design and analysis of the 3-DTEL in parallel cluster system. Cluster applications have more processor cores to manage and exploit the computational capacity of high-end machines providing effective and efficient means of parallelism even as the challenges of providing effective resources management grows. It is a known fact that high capacity computing platform are expensive, and are characterized by long-running, high processor-count jobs. The performance of message-passing programs depends on the parallel target machine, and the parallel programming model to be applied to achieve parallelism. In a cluster machines having large number of processing units' scalability becomes an important issue. Many programs from scientific computing have a large potential for parallelism that is exploited best in such a programming model for mixed fast and data parallelism where the parallelism can be structured in the form of concurrent multi-processor tasks [21].

Developing parallel applications have its own challenges in the field of parallel computing. With reference to [11], there are theoretical challenges such as task decomposition, dependence analysis, and task scheduling. Then they are practical challenges such as portability, synchronization, and debugging. However, there exist an alternative and cost effective way of achieving performance through the use of loosely connected system of processors with a local area network [3]. Hence, for a global task with other processors relevant data needs to be passed from processor to processor through a message-passing mechanism [20, 15], since there is greater demand for computational speed and the computations must be completed within reasonable time period. A multi-processor task can be implemented on a subset of processors, and one of the advantages is based on the fact that for many message-passing machines communication costs are affected by the number of participating processors.

Design and analysis for finite difference DD for 2-D heat equation has been discussed in [23], and the parallelization for 3-DTEL on parallel virtual machine with DD [8] show effective load scheduling over various mesh sizes, which produce the expected inherent speedups. Parallel algorithms have been implemented for the finite difference method by [12], and [21, 13] use the discrete eigen functions method with the AGE method on telegraph equation problem.

The theoretical properties of the 3-D ADI algorithm with the parallel design approach employing SPMD model with DD are promising, achieving good performance as to what was done by [7] in practice can be challenging. There is a tradeoff between the reduction of the time required for an inherently sequential part of an algorithm, and an increase in the number of the iterations required to converge [2]. Previous work on 3-D ADI scheme did not consider the parallel design approach on parallelism and improvement on scalability. To write SPMD programs using one of the standard message-passing software like MPI [13] requires the explicit administration of processors with a large user group. In this paper, we present a support for the implementation of parallel design and analysis with the use of DD strategy. Our programming style allows the application programmer to specify the program organization in a clear and readable program code.

We presented a detailed study of using parallel design and analysis on 3-DTEL, and solved by the use of ADI method on a GCC cluster MPI. The SPMD model is employed with DD to enhance overlapping communication with computation that resulted in significant improved speedup, effectiveness, and efficiency across varying mesh sizes as compared to [7].

Our results demonstrated the overlap communication with computation, and the ability to arbitrary use of varying mesh sizes distribution on GCC to reduce memory pressure while preserving parallel efficiency. On the other hand, the advantage of our platform is to have somewhat specification mechanism through a static distribution, and an execution implementation.

The rest of the paper is organized as follows. Section 2 presents related work. Section 3 introduces the model for the 3-DTEL and the 3-D ADI scheme. Section 4 introduces the parallel design and analysis. Section 5 introduces the results of several experiments, which illustrate and evaluate the parallelization possible with our platform. Section 6 gives the conclusion.

II. Related Work

A work by [16] achieved configuration of MPI-based message passing programs, and various other platforms for the application of telegraph and heat equations have been done in [7, 8]. Description of application aware job scheduler that dynamically controls resource allocation among concurrently executing jobs was done by [22]. A framework called 'Gridway' for adaptive execution of applications in Grids was described by [14]. Parallelization by time decomposition was first proposed by [18] with motivation to achieve parallel real-time solutions, and even the importance of loop parallelism, loop scheduling have been extensively studied [1]. The ADI method for the Partial Differential Equations (PDE) proposed by [19] has been widely used for solving algebraic systems resulting from finite difference method analysis of PDE in several scientific and engineering applications. Works on parallel implementation of 2-D Telegraph problem on cluster systems have been done in [10, 12].

In [12] the unconditional stability of the alternating difference schemes has similarity to our scheme and shows that the unconditional stability application is useful to its speedup and efficiency as studied. Our implementation in the GCC platform has several aspects that differentiate it from the above. GCC is designed for application running on distributed memory clusters, which can dynamically and statically calculate partition sizes based on the run-time performance of the application. We use an efficient algorithm with stability which maps data using message passing over the GCC cluster. We evaluated our system using experimental results from speedup and efficiency for the system utilization. Our approach is best suited to applications where data and computations are uniformly distributed across processors.

III. The Model Problem

We consider the second order telegraph equation in 3-D:

$$\frac{\partial^2 v}{\partial t^2} + a\frac{\partial v}{\partial t} - \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2}\right) = 0 \qquad (3.1)$$

where $a = RC + GL$, let $\Delta x, \Delta y \; and \; \Delta z$ be the grid spacing in the x, y, z and t directions, where $\Delta x = \Delta y = \Delta z = 1/m$, m is a positive integer. We can solve (3.1) by extending the 1-D simple implicit finite difference method [21] of the telegraph equation to the above 3-D telegraph equation, (3.1) becomes:

$$\frac{v_{i,j,k}^{n+1} - 2v_{i,j,k}^{n} + v_{i,j,k}^{n-1}}{(\Delta t)^2} + a\frac{v_{i,j,k}^{n+1} - v_{i,j,k}^{n-1}}{2\Delta t} - \left\{ \begin{array}{l} \dfrac{v_{i+1,j,k}^{n+1} - 2v_{i,j,k}^{n+1} + v_{i-1,j,k}^{n+1}}{(\Delta x)^2} \\ + \dfrac{v_{i,j+1,k}^{n+1} - 2v_{i,j,k}^{n+1} + v_{i,j-1,k}^{n+1}}{(\Delta y)^2} \\ + \dfrac{v_{i,j,k+1}^{n+1} - 2v_{i,j,k}^{n+1} + v_{i,j,k-1}^{n+1}}{(\Delta z)^2} \end{array} \right\} = 0 \qquad (3.2)$$

although this simple implicit scheme is unconditionally stable, therefore, the computational time is extremely huge.

A. ADI Method on 3-DTEL

We derive the ADI method for 3-DTEL of the simple implicit finite difference method by using a general ADI procedure [6] extended to (3.1). The ADI method is a well-known method for solving the PDE. The main feature of ADI is to sweep directions alternatively. In contrast to the standard finite-difference formulation with only one iteration to advance from the *nth* to (*n + 1*)*th* time step, the formulation of the ADI method requires multilevel intermediate steps to advance from the *nth* to (*n + 1*)*th* time step. Equation (3.2) can be rewritten as:

$$\left(I + \sum_{m=1}^{3} A_m\right) v_{i,j,k}^{n+1} - 2C_o v_{i,j,k}^{n} + C_1 v_{i,j,k}^{n-1} = 0 \qquad (3.3)$$

where the operators of I, $A_m s$, and the constants of C_o, C_1 are define as:

$$Iv_{i,j,k}^{n} \equiv v_{i,j,k}^{n} \qquad (3.4)$$

$$A_1 v_{i,j,k}^{n} \equiv -\rho_x \left(v_{i+1,j,k}^{n} - 2v_{i,j,k}^{n} + v_{i-1,j,k}^{n}\right) \qquad (3.5)$$

$$A_2 v_{i,j,k}^{n} \equiv -\rho_y \left(v_{i,j+1,k}^{n} - 2v_{i,j,k}^{n} + v_{i,j-1,k}^{n}\right) \qquad (3.6)$$

$$A_3 v_{i,j,k}^{n} \equiv -\rho_z \left(v_{i,j,k+1}^{n} - 2v_{i,j,k}^{n} + v_{i,j,k-1}^{n}\right) \qquad (3.7)$$

$$C_o \equiv \frac{1}{(\Delta t)^2} \Big/ \left(\frac{1}{(\Delta t)^2} + \frac{a}{2\Delta t}\right) \qquad (3.8)$$

$$C_1 \equiv \left(\frac{1}{(\Delta t)^2} - \frac{a}{2\Delta t}\right) \Big/ \left(\frac{1}{(\Delta t)^2} + \frac{a}{2\Delta t}\right) \qquad (3.9)$$

the constant of ρ_x, ρ_y and ρ_z are:

$$\rho_x = \frac{b}{(\Delta x)^2} \Big/ \left(\frac{1}{(\Delta t)^2} + \frac{a}{2\Delta t}\right) \qquad (3.10)$$

$$\rho_y = \frac{b}{(\Delta y)^2} \Big/ \left(\frac{1}{(\Delta t)^2} + \frac{a}{2\Delta t} \right) \tag{3.11}$$

$$\rho_y = \frac{b}{(\Delta y)^2} \Big/ \left(\frac{1}{(\Delta t)^2} + \frac{a}{2\Delta t} \right) \tag{3.12}$$

TABLE I. THE ADI 3-DTELALGORITHM

The ADI 3-DTEL Algorithm

Input = $v_{i,j,k}^{n}, v_{i,j,k}^{n-1} \ \forall i, j, k$

Output = $v_{i,j,k}^{n+1} \ \forall i, j, k$

Begin

Sub-Iteration 1:

$$-\rho_x v_{i+1,j,k}^{n+1(1)} + (1+2\rho_x) v_{i,j,k}^{n+1(1)} - \rho_x v_{i-1,j}^{n+1(1)} =$$
$$-(A_2 + A_3) v_{i,j,k}^{n+1(*)} + (2C_o v_{i,j,k}^{n} - C_1 v_{i,j,k}^{n-1})$$
$$\forall i, j, k$$

Sub-Iteration 2:

$$-\rho_y v_{i,j+1,k}^{n+1(2)} + (1+2\rho_y) v_{i,j,k}^{n+1(2)} - \rho_y v_{i,j-1,k}^{n+1(2)}$$
$$= v_{i,j,k}^{n+1(1)} + A_2 v_{i,j,k}^{n+1(*)}. \quad \forall i, j, k$$

Sub-Iteration 3:

$$-\rho_z v_{i,j,k+1}^{n+1(3)} + (1+2\rho_z) v_{i,j,k}^{n+1(3)} - \rho_z v_{i,j,k-1}^{n+1(3)}$$
$$= v_{i,j,k}^{n+1(2)} + A_3 v_{i,j,k}^{n+1(*)}. \quad \forall i, j, k$$

End

and set

$$v_{i,j,k}^{n+1(*)} = 2v_{i,j,k}^{n} - v_{i,j,k}^{n-1} \tag{3.13}$$

which is a prediction of $v_{i,j,k}^{n+1}$ by the extrapolation method.

Then splitting (3.3) by using an ADI procedure as in [17], we get a set of recursion relations as follows:

$$(I + A_1) v_{i,j,k}^{n+1(1)} = -(A_2 + A_3) v_{i,j,k}^{n+1(*)} + (2C_0 v_{i,j,k}^{n} - C_1 v_{i,j,k}^{n-1}) \tag{3.14}$$

$$(I + A_2) v_{i,j,k}^{n+1(2)} = v_{i,j,k}^{n+1(1)} + A_2 v_{i,j,k}^{n+1(*)} \tag{3.15}$$

$$(I + A_3) v_{i,j,k}^{n+1(3)} = v_{i,j,k}^{n+1(2)} + A_3 v_{i,j,k}^{n+1(*)} \tag{3.16}$$

where $v_{i,j,k}^{n+1(1)}, v_{i,j,k}^{n+1(2)}$ are the intermediate solutions and the desired solution is $v_{i,j,k}^{n+1} = v_{i,j,k}^{n+1(3)}$. Finally, expanding A_1, A_2 and A_3 on the left side of (3.14) and (3.16), we get the 3-D ADI algorithm as in Table 1.

IV. PARALLEL IMPLEMENTATION, DESIGN AND ANALYSIS

A. The Parallel Platform

The Geranium Cadcam Cluster consist of 32 Intel Pentium dual core processor at 1.73GHZ and 0.99GB RAM. Communication is through a fast Ethernet of 100 MBits per seconds running Linux, located at the University of Malaya. The cluster performance has high memory bandwidth with a message passing supported by MPI [13]. The program is written in C and provides access to MPI through calling MPI library routines. The platform contains more computations on varying set of mesh sizes. Performance in the platform concerns the resource assessment and code placement on computing resources [5]. The 3-DTEL with ADI scheme is implemented on the GCC cluster, with an objective to evaluate the overhead it introduces with ability to exploit the inherent parallelism of the computation. We observed the scalability across the varying number of processors and mesh sizes, to enable the speedup we need convergence in fewer than N iterations.

B. Domain Decomposition

The parallelization of the computations is implemented by means of grid partitioning technique. The computing domain is decomposed into many blocks with reasonable geometries. Along the block interfaces, auxiliary control volumes containing the corresponding boundary values of the neighboring block are introduced, so that the grids of neighboring blocks are overlapped at the boundary. When the domain is split, each block is given an I-D number by a "master" task, which assigns these sub-domains to "slave" tasks running in individual processors. In order to couple the sub-domains' calculations, the boundary data of neighboring blocks have to be interchanged after each iteration. The calculations in the sub-domains use the old values at the sub-domains' boundaries as boundary conditions. This may affect the convergence rate; however, because the algorithm is implicit, the blocks strategy can preserve nearly same accuracy as the sequential program.

The DD is used to distribute data between different processors; the static load balancing is used to maintain same computational points for each processor. The partitioning and load balancing is done in the pre-processing stage giving no room for extra storage when the parallel program is executed. Data parallelism originated the SPMD [17], thus, the finite difference approximation used in this paper can be treated as an SPMD problem. Same computation is performed for multiple data sets, and the multiple data are different parts of the overall grid.

C. Parallel ADI with MPI

We focus on computational domain partitions in implementing the parallel 3-DTEL ADI scheme on GCC platform. We need divide the dimensions into sub-domains with no unique way of partitioning the domain of computation. The case of making a balance between the implementation of the algorithm and the communication efficiency is paramount to balance. The partitioning considered is the orientation of slices changing with the sweeps according to [4].

After *x*-sweeps, the orientation changes to the *y* or the *z* direction. In this process each processor owns three data domains, one for each direction. Implementing the parallel algorithm for solving (3.1) is based on: indication of sweeping direction for each sub-domain. Sweeping direction of each sub-domain must be in opposite direction of its neighbors. For example, we must use left right direction for odd sub-domains and right left direction for even sub-domains. Updating start node of each sub-domain with (3.14) and (3.16), each processor of the parallel machine works only on its specific portion of the grid and when processor needs information from the nearest neighbor a message is passed through the MPI message passing library. For the best parallel performance, one would like to have optimal load balancing and as little communication between processors as possible. Considering load balancing first, one would like each processor to do exactly the same amount of work, hence, each processor is not idle. For the finite difference code, the basic computational element usually is the node; it makes sense to partition the grid such that each processor gets an equal number of nodes to work on. The second criterion is that the amount of communication between processors be made as small as possible. To minimize communication, the program must divide the domain in a way that minimizes the length of the touching faces in the different sub-domains. The number of processors that one processor has to communicate with also contributes to additional communication time, because of the latency penalty for starting the new message. At first step, we divide the spatial computational domain to $P = P_1 \times P_2 \times P_3$. We can use the non-blocking message passing for this communication stage to reduce computing time by allowing work to be done while communication is in progress.

D. Load Balancing

With static load balancing, the computation time of parallel subtasks should be relatively uniform across processors; otherwise, some processors will be idle waiting for others to finish their subtasks. Therefore, the domain decomposition should be reasonably uniform. A better load balancing is achieved with the pool of tasks strategy, which is often used in master – slave programming [2]: the master task keeps track of idle slaves in the distributed pool and sends out the next task to the first available idle slave. With this strategy, the processors are kept busy until there is no further task in the pool. If the tasks vary in complexity, the most complex tasks are sent out to the most powerful processor first. With this strategy, the number of sub-domains should be relatively large compared to the number of processors.

Otherwise, the slave solving the last sent block will force others to wait for the completion of this task; this is especially true if this processor happens to be the least powerful in the distributed system. The block size should not be too small either, since the overlap of nodes at the interfaces of the sub-domains become significant. This results in a doubling of the computations of some variables on the interfacial nodes, leading to a reduced efficiency. Increasing the block number also lengthens the execution time of the master program, which leads to a reduced efficiency.

E. Speedup and Efficiency

A simple speedup analysis with reference to [2] produces the following:

$$\varphi = \frac{N}{Nr(K+1)+K}, \tag{4.1}$$

where r is the ration of the time taken by coarse propagation to fine propagation over the same time interval, K is the number of iterations required for convergence, and communication overhead is ignored. In the limit $r \to 0,\ \varphi \to \frac{N}{K}$, therefore, the efficiency will be $\frac{1}{K}$. The algorithm for the scheme is performed on a distributed memory system of p processors, assumes that each processors initially stores $n = N/p$ objects distributed over the entire physical domain. In the first iteration of the algorithm, the domain is decomposed into two sub-domains so that the difference between the sums of the weight of the sub-domain is as small as possible. Then the same process is applied to two sub-domains in parallel, and process is repeated recursively, for log p iteration. In other words, during iteration i, $1 \le i \le \log p$, the p processors are group into 2^{i-1} groups of $p/2^{i-1}$ processors each. At the beginning of the iteration, the problem domain is already partitioned into 2^{i-1} sub-domains and the objects in each sub-domain are stored in single group of processors. At the end of the iteration, each processor group is divided into two groups, and the corresponding sub-domain is divided into two sub-groups with the object in one sub-domain residing in one half the processors and the other objects in the other sub-domain residing in the other half of processor

V. Results and Discussion

Consider the Telegraph Equation of the form:

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial v}{\partial z^2} = \frac{\partial^2 v}{\partial t^2} + \frac{\partial v}{\partial t} + v \tag{5.1}$$

the boundary condition and initial condition posed are:

$$\left.\begin{aligned} v(0,y,z) &= 0 \\ v(1,y,z) &= 100 \\ v(x,0,z) &= 0 \\ v(x,1,z) &= 100 \\ v(x,y,0) &= 0 \\ v(x,y,1) &= 100 \end{aligned}\right\} \quad t \ge 0 \tag{5.2}$$

$$v(x,y,z) = e^{xyz}, \tag{5.3}$$

A. Parallel Efficiency

The speedup and efficiency obtained for various sizes, for 70x70x6 to 210x210x6, are for various numbers of sub-domains, for B = 50 are listed in Tables 2 – 4. In the Tables

we also listed the wall (elapsed) time for the master task, T_W, (this is necessarily greater than the maximum wall time returned by the slaves), the master CPU time, T_M, the average slave computational time, T_{SC}, and the average slave data communication time, T_{SD}, all in seconds. The speedup and efficiency versus the number of processors are shown in Fig. 1 and Fig. 2, respectively, with block number B as a parameter.

The results in the Tables show that the parallel efficiency increases with increasing grid size for given block number, and decreases with the increasing block number for given grid size. As the number of processors increase, though this leads to a decrease in execution time, but a point is reached when the increased processors will not have much impact on total execution time. Hence, when the numbers of processors increase, balancing the number of computational cells per processors will become a difficult task due to significant load imbalance. The gain in increasing execution time for certain mess sizes is due to uneven distribution of the computational cell, and the execution time has a very small change due to DD influence on performance in parallel computation.

The total CPU time is composed of three parts: the CPU time for the master task, the average slave CPU time for data communication and the average slave CPU time for computation, $T = T_M + T_{SD} + T_{SC}$.

TABLE II. THE WALL TIME T_W, THE MASTER TIME T_M, THE SLAVE DATA TIME T_{SD}, THE SLAVE COMPUTATIONAL TIME T_{SC}, THE TOTAL TIME T, THE PARALLEL SPEED-UP S_{PAR} AND THE EFFICIENCY E_{PAR} FOR A MESH OF 70X70X6, WITH B = 50 BLOCKS AND $NITER$ = 100.

N	T_w	T_m	T_{sd}	T_{sc}	T	S_{par}	E_{par}
1	1245	38	4	522	564	1.000	1.000
2	621	36	3	281	320	1.761	0.881
3	318	36	3	188	227	2.482	0.827
4	257	36	3	158	197	2.865	0.716
5	238	36	3	131	170	3.324	0.665
6	219	36	3	107	146	3.864	0.644
7	206	36	3	92	131	4.321	0.617
8	205	36	3	76	115	4.918	0.615
12	183	36	3	60	96	5.921	0.493
16	176	36	3	44	83	6.824	0.427
20	155	36	3	28	67	8.211	0.411
24	138	36	3	25	64	8.926	0.372
28	125	36	3	21	60	9.412	0.336
32	112	36	3	14	53	10.896	0.341

TABLE III. THE WALL TIME T_W, THE MASTER TIME T_M, THE SLAVE DATA TIME T_{SD}, THE SLAVE COMPUTATIONAL TIME T_{SC}, THE TOTAL TIME T, THE PARALLEL SPEED-UP S_{PAR} AND THE EFFICIENCY E_{PAR} FOR A MESH OF 120X120X6, WITH B = 50 BLOCKS AND $NITER$ = 100.

N	T_w	T_m	T_{sd}	T_{sc}	T	S_{par}	E_{par}
1	2721	119	13	1589	1721	1.000	1.000
2	1292	113	13	822	948	1.818	0.909
3	694	113	13	497	623	2.764	0.921
4	482	113	13	347	473	3.641	0.910
5	449	113	13	325	451	3.817	0.763
6	408	113	13	243	369	4.663	0.777
7	396	113	13	225	351	4.912	0.702
8	385	113	13	167	293	5.873	0.734
12	371	113	13	135	261	6.618	0.552
16	372	113	13	97	223	7.738	0.484
20	348	113	13	59	185	9.328	0.466
24	322	113	13	37	163	10.611	0.442
28	308	113	13	28	154	11.322	0.404
32	284	113	13	12	138	12.589	0.393

TABLE IV. THE WALL TIME T_W, THE MASTER TIME T_M, THE SLAVE DATA TIME T_{SD}, THE SLAVE COMPUTATIONAL TIME T_{SC}, THE TOTAL TIME T, THE PARALLEL SPEED-UP S_{PAR} AND THE EFFICIENCY E_{PAR} FOR A MESH OF 210X210X6, WITH B = 50 BLOCKS AND $NITER$ = 100.

N	T_w	T_m	T_{sd}	T_{sc}	T	S_{par}	E_{par}
1	13825	378	55	8086	8519	1.000	1.000
2	6439	374	54	4189	4617.3	1.845	0.923
3	3427	374	54	2662	3090	2.757	0.919
4	2718	374	54	1909	2337	3.646	0.914
5	2589	373	54	1548	1975	4.315	0.863
6	2443	373	54	1286	1713	4.974	0.829
7	2094	373	54	1124	1551	5.495	0.785
8	2019	373	54	970	1398	6.184	0.773
12	1924	373	54	562	989	8.616	0.718
16	1918	373	54	396	823	10.352	0.647
20	1710	373	54	278	705	12.1	0.605
24	1621	373	54	230	656.6	12.984	0.541
28	1597	373	54	163	591	14.448	0.516
32	1481	373	54	132	558	15.264	0.477

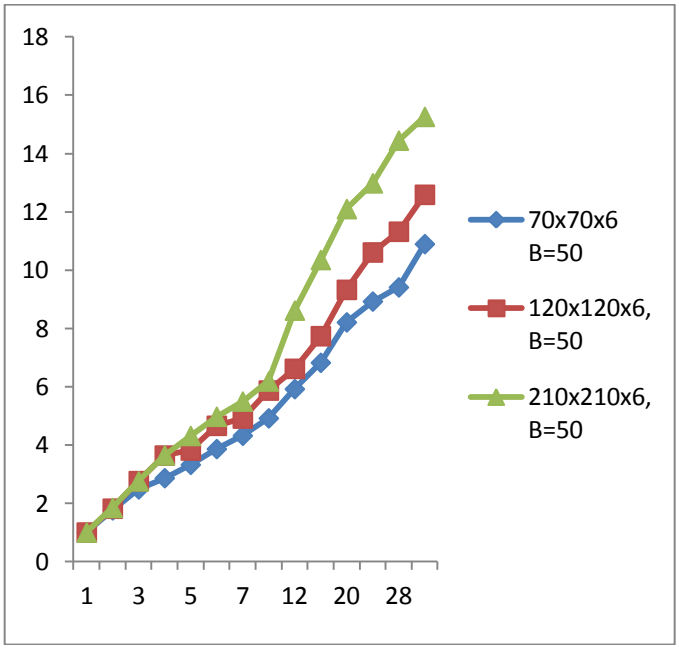


Fig. 1. Speedup versus the number of processors for mesh 70x70x6,120x120x6 and 210x210x6

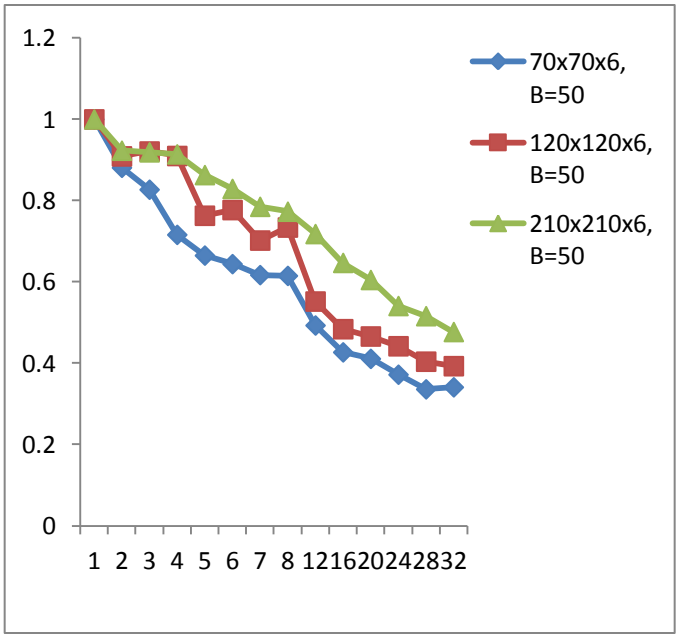


Fig. 2. Parallel efficiency versus the number of processors for mesh 70x70x6, 120x120x6 and 210x210x6

B. Numerical Efficiency

The numerical efficiency includes the DD efficiency and convergence rate behavior. The DD efficiency includes the increase of floating point operations induced by grid overlap at interfaces and the CPU time variation generated by DD techniques. In Table 5, we listed the total CPU time distribution over various grid sizes and block numbers running with only one processor. In Table, the DD efficiency can be calculated, and the result as shown in Fig. 3. Note that the DD efficiency can be greater than one, even with one processor. Fig. 3 also shows that the optimum number of sub-domains, which maximizes the DD efficiency E_{DD}, increases with the grid size. The convergence rate behavior, the ratio of the iteration number for the best sequential CPU time on one processor and the iteration number for the parallel CPU time on n processor, describes the increase in the number of iterations required by the parallel method to achieve a specified accuracy, as compared to the serial method. This increase is caused mainly by the deterioration in the rate of convergence with increasing number of processors and sub-domains. Because the best serial algorithm is not known generally, we take the existing parallel program running on one processor to replace it. Now the problem is that how the decomposition strategy affects the convergence rate? The results are summarized in Table 6 and Fig. 4, and Table 7 and Fig. 5.

It can be seen that the convergence rate decreases with increasing block number and increasing number of processors for given grid size. The larger the grid size, the higher the convergence rate.

TABLE V. THE TOTAL COMPUTATIONAL TIME *T* FOR 100 ITERATIONS AS A FUNCTION OF VARIOUS BLOCK NUMBERS

	B = 1	*B = 8*	*B = 16*	*B = 24*	*B = 50*
70x70x6	*411*	*437*	*481*	*509*	*564*
120x120x6	*572*	*641*	*987*	*1394*	*1721*
210x210x6	*3493*	*4168*	*4928*	*6294*	*8519*

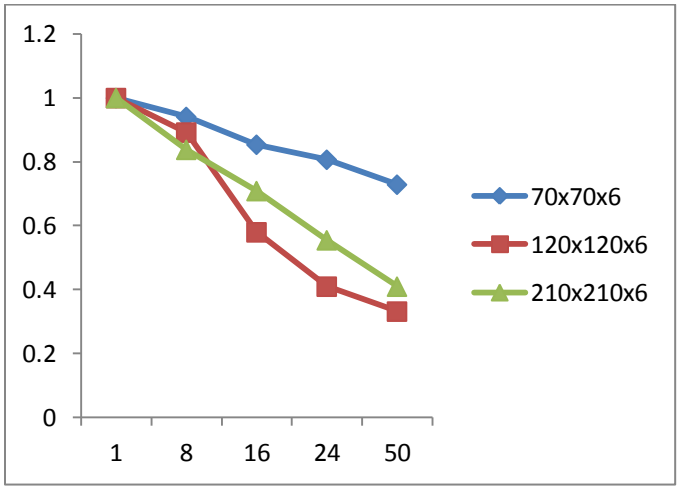


Fig. 3. The DD efficiency versus the number of sub-domains for various meshes.

TABLE VI. THE NUMBER OF ITERATION TO ACHIEVE A GIVEN TOLERANCE OF 10^{-3} FOR A GRID OF 70X70X6

N	*B = 1*	*B = 16*	*B = 50*
1	*1796*	*1987*	*2129*
2	*1796*	*2206*	*2346*
4	*1796*	*2293*	*2492*
8	*1796*	*2371*	*2524*

12	*1796*	*2396*	*2598*
16	*1796*	*2417*	*2609*
28	*1796*	*3214*	*4968*
32	*1796*	*3486*	*5291*

TABLE VII. THE NUMBER OF ITERATION TO ACHIEVE A GIVEN TOLERANCE OF 10^{-2} FOR A GRID OF 120X120X6

N	*B = 1*	*B = 16*	*B = 50*
1	*2138*	*2313*	*2434*
2	*2138*	*2329*	*2518*
4	*2138*	*2348*	*2531*
8	*2138*	*2461*	*2687*
12	*2138*	*2461*	*2692*
16	*2138*	*2518*	*2698*
28	*2138*	*3763*	*5321*
32	*2138*	*3775*	*5711*

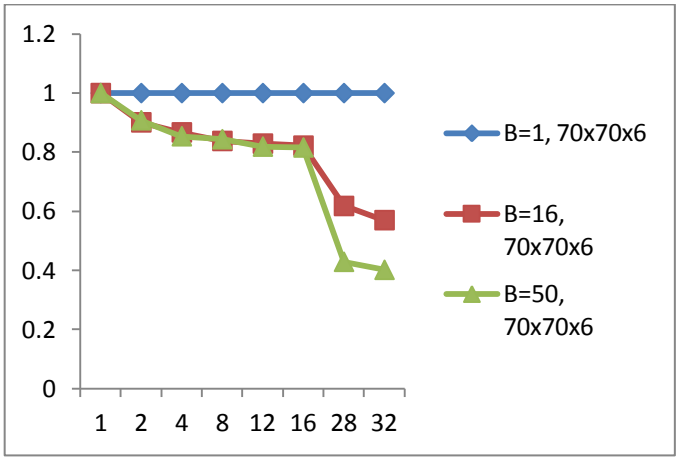


Fig. 4. Convergence behavior with domain decomposition for mesh 70x70x6

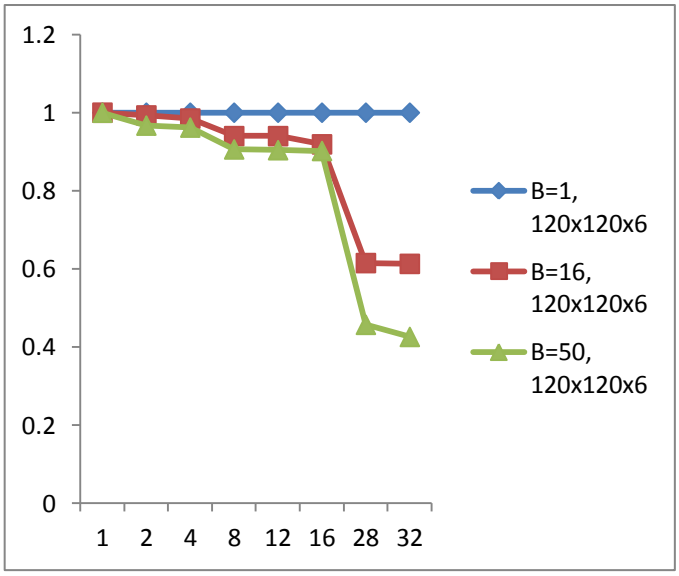


Fig. 5. Convergence behavior with domain decomposition for mesh 120x120x6

VI. CONCLUSION

The results presented in this paper show the study on the parallel design and analysis for 3-D TEL ADI scheme with MPI. The objective is to present a design for the GCC for distributed computation, because they depend on empirical concern. The system allows a parallel collection of overlapping communication to avoid unnecessary synchronization and to have the impact of parallel convergence. In addition to the use of ease of our platform, compared to other approaches show negligible overhead with effective load scheduling over various mesh sizes, which produce the expected inherent speedups. It was also confirmed that flexible scheduling for the overlapping communication are important, and this is easy on with SPMD model as seen from the Tables and Figures. Computational results obtained have clearly shown the benefits of parallelization. The DD greatly influences the performance of the 3-DTEL ADI scheme on the parallel computers. On the basis of the current parallelization strategy, more sophisticated models can be attacked efficiently. Similarly, we are interested in improving our algorithms and testing implementations on additional architectures.

VII. FUTURE WORK

The description of 3-DTEL with the use of ADI method on GCC Cluster System with MPI employing the SPMD technique has been carried out. This paper allows a parallel collection of overlapping communication to avoid unnecessary synchronization and to have the impact of parallel convergence. We suggest future work to be carried out on the 3-DTEL employing the used of Iterative Alternating Direction Implicit (IADE) method. Parallel implementation for the scheme could use the Input File Affinity Measure on a tightly coupled distributed environment with dynamic allocation of task with varying mesh sizes.

REFERENCES

[1] J. Aguilar, E. Leiss, 'Parallel Loop Scheduling Approaches for Distributed and Shared Memory System', Parallel Process Letter 15 (1 – 2), 2005, pp. 131 – 152

[2] E. Aubanel, 'Scheduling of tasks in the parareal algorithm' Parallel Computing 37 (3), 2011, 172 – 182

[3] W. Barry, A. Michael, '*Parallel Programming Techniques and Application using Networked Workstation and Parallel Computers*' 2003, Prentice Hall, New Jersy

[4] G. Baolai. On the Performance of Parallel Implementation of an ADI Scheme for Parabolic PDEs on Shared and Distributed Memory. Shared Hierarchical Research Computing Network, The University of Western Ontario.

[5] D. Cyril, M. Fabrice, 'Jacobi computation using mobile agent' Int'l Journal of Computer Science & Information Technologies, 1 (5), 2010, 392 – 401

[6] D.J Evans, B. Hassan, 'Numerical Solution of the Telegraph Equation by the AGE Method', Int'l Journal of Computer Mathematics Vol. 80, number 10, 2003, pp 1289 – 1297

[7] S. U. Ewedafe, H. S. Rio, 'Parallelization of 2-D IADE-DY Scheme on Geranium Cadcam Cluster for Heat Equation' Int'l Jour. Of Advanced Research in Artificial Intelligence, 2 (6), 2013, pp. 27 – 33

[8] S. U. Ewedafe, H. S. Rio, 'Parallelization of 3-D ADI Scheme on Telegraph Problem using Domain Decomposition with PVM' Int'l Jour. Of Applied Information Systems, 4 (11), 2012, pp. 12 – 24

[9] S. U. Ewedafe, H. S. Rio, 'Parallel Implementation of 2-D Telegraph Equation on MPI/PVM Cluster' Int'l Jour. of Parallel Programming, 39, Issue 2, 2011, 202 – 231

[10] S. U. Ewedafe, H. S. Rio, 'Armadillo Generation Distributed Systems & Geranium Cadcam Cluster for solving 2-D Telegraph Equation' Int'l Jour. of Computer Mathematics, 88, Issue 3, 2011, 589 – 609

[11] N. Giacaman, O. Sinnen, 'Parallel iterator for parallelizing object-oriented applications' Intl journal of parallel programming, 39 (2), 2011, 223 – 269, 2011.

[12] Y. Guang-Wei, Long-Jun S., Yu-Lin Z., 'Unconditional Stability of Parallel Alternating Difference Schemes for Semilinear parabolic Systems' Applied Mathematics and Computation 117, 2001, pp 267 – 283

[13] W. Groop, E. Lusk, A. Skjellum, 'Using MPI, portable and parallel programming with the message passing interface,' 1999, 2nd Ed., Cambridge MA, MIT Press

[14] E. Huedo, R. Montero, I. Llorente, 'A Framework for Adaptive Execution in Grids' Software Practice & Experiences 34 (7), 2004, pp. 631 - 651

[15] K. Jaris, D.G. Alan, 'A High-Performance Communication Service for Parallel Computing on Distributed System', Parallel Computing 29, 2003, pp 851 – 878

[16] L, Kale, S. Kumar, J. DeSouza, 'A malleable-Job System for Time-Shared Parallel Machines' Proceedings of the second IEEE/ACM Int'l Symposium on Cluster Computing, IEEE Computer Society, 2002, Washington DC, U.S.A,

[17] H. Laurant, 'A method for automatic placement of communications in SPMD parallelization' Parallel computing 27, 2001, 1655 – 1664

[18] J. L. Lions., Y. Maday, G. Turinki, 'Parareal in time discretization of PDE' Comptes, rendus de lacadimie des sciences – series 1 – mathematics 332 (7), 2011, 661 – 668

[19] D.W Peaceman, H.H Rachford, 'The Numerical Solution of Parabolic and Elliptic Differential Equations' Journal of Soc. Indust. Applied Math. 8 (1), 1955, pp 28 – 41

[20] Peizong L., Z. Kedem, 'Automatic Data and Computation Decomposition on Distributed Memory Parallel Computers' ACM Transactions on Programming Languages and Systems, vol. 24, number 1, 2002, pp 1 – 50

[21] T. Rauber, G. Runger, 'A Transformation Approach to Derive Efficient Parallel Implementations', IEEE Transactions on Software Engineering, 26 (4), 2000, pp. 315 - 399

[22] J. Weissman, L. Rao, D. England, 'Integrated Scheduling: The Best of both Worlds', Jour. Parallel & Distri. Computing 63 (6), 2003, pp. 631 - 651

[23] W. Zheng-Su, Z. Baolin, C. Guang-Nan, 'Design and analysis for finite difference DD for 2-D heat equation', ICA3PP – 02, IEEE Computer Society

Estimating Traffic Intensity at Toll Gates Using Queueing Networks

Vincent O. R
Department of Computer Science
Federal University of Agriculture
Abeokuta, Nigeria

Olayiwola O. E.
Department of Computer Science
Federal University of Agriculture
Abeokuta, Nigeria

Kosemani O. O.
Department of Computer Science
Federal University of Agriculture
Abeokuta, Nigeria

***Abstract*—Traffic information generation is a routine-like operation that is done on a daily basis at any public gate. A toll gate is a public roadway by which people enter and leave a public organisation. The existing models give premium consideration to security over prompt services and as such associated with processes that have high cost of implementation, inaccuracy from complex method as well pose other technical problems such as delay. This research presents an automated procedure for monitoring traffic at toll gates to give the best compromise among the conflicting objectives of payment, security and good services. The system gathers information about the traffic situation with respect to the license plate number captured from each vehicle that passes through the toll gate and as well captures data such as arrival speed, arrival time and date and uses this data as input to generate traffic report/information on a daily basis. Experimentally the system shows that it can effectively capture the vehicle video and detect the license plate in day time, showing accuracy of about 85% to 90%, practical results based on actual data are included.**

General Terms: Image processing; Artificial intelligence; Information Engineering

Keywords—Toll Gates; Queueing Networks; Traffic intensity; Delay and Image detection

I. INTRODUCTION

A toll gate is a point of entry to a space enclosed by walls, or a moderately sized opening in some sort of fence. Gates may prevent or control the entry or exit of individuals, or they may be merely decorative. The main advantage of a toll gate is the opportunity to keep track of vehicles plying the highways, bridges, and tunnels on which the system is installed. The system enhances the collection of entrance payments which is done either manually or electronically. The manual mode of payment is considered primitive because it poses problems such as congestion at toll gates, especially during festive seasons when traffic tends to be heavier than normal [1]. This incontinency results in fatigue and inaccuracy in the automation of the system.

On the other hand, the electronic entrance system executes automatic payments using wireless communication without long stoppings at the express way. The problems with electronic system are the reflected, distracted, shadowed waves experienced by canopy, wall, or, booth and interference waves from neighboring lanes which may constitute threats to the operations of the system [2].

With the increasing number of vehicles on roads, traffic at gates could no longer be estimated manually but rather electronically. Electronic toll gates also help in enforcing laws and traffic rules for smooth flow of it. The number of vehicles moving in and out at different speeds and the timing are basic data that could generate information about the traffic situation. Entrance-gates are located majorly on freeways and parking structures for general checking purposes most importantly for payments and security. Toll gates are enabled with traffic management systems to check for vehicles moving at higher speeds. The purposes of creating toll gates were achieved at the expense of vehicle owners. The important question which has over the years not been answered is what are the implications of toll gate on vehicle owners and the society at large?

The literatures on toll gates are in several categories; research works either extract the license plate of vehicles for tracking or for estimating vehicles speed. Many papers analysed low frame-rate video taken from an un-calibrated camera and estimate mean traffic speed for tracking of vehicles [3]. Others dwelt on plate detection. A research presented an algorithm that allows automatic ID measurements and subsequent estimation of vehicle speed from single un-calibrated images [4]. Some use canny edge detection to detect vehicles speed [5], others used morphological operators to detect vehicle number plate from the videos of the vehicles in different illumination environments, but the central idea of license plate extraction is through video [6]. Another developed traffic monitoring using vehicle-based sensors of taxi with two traffic status estimation algorithms adopted: link-based and the vehicle based on sparse and incomplete information [7]. However, the sensors were always set with long sampling interval because of communication cost saving and network congestion avoidance.

Xie proposed a Privacy Awareness Monitoring System (PAMS) which works as aggregate query processor to protect the location privacy of drivers as it anonymizes the IDs of vehicles [8]. Though PAMS answers the problems of high queries and accuracy and also achieves good balances among privacy, accuracy and efficiency for traffic monitoring, purpose-centred mission were observed which were always aimed to achieve security over other inconveniences.

This work presents a system for license plate extraction from a video frame. It focuses on extracting general information of a vehicle arriving at toll gate, information such

as its arrival speed and license plate number of vehicle and these information combined to estimate the traffic intensity at the gate given a sequence of real-time traffic videos that is taken from a camera strategically installed in the toll gate environment to capture the video of the activities of the booth when the vehicles pass through and generated images from which license plate number are extracted and then processed. The work solves the problem of inaccuracy and fatigue resulting from manually generating traffic information at toll gates as well as high cost of installation and maintenance of existing automated systems. The motivation of the work is to find a cost effective method to detect the vehicles' arriving speed, and their identification number, and using all these data as information to estimate the traffic intensity at the booth. The paper is arranged as follows: Section 2 presents the review of literature while section 3 describes the traffic monitoring system with queueing theory. Section 4 gives the results and evaluation, Section 5 concludes the work.

II. Related Methods

A. Automated Toll Gate System

Automated Toll gate System is considered the most sophisticated entrance roads in the world [9]. Cameras are equipped with Optical Character Recognition (OCR). The OCR cameras are used to capture license plate numbers of vehicles without transponders. The entrance bill is then sent directly to the registered address of the vehicle owners. There are two laser beam scanners in the system which is placed above the roadway to detect the types of vehicles passing through the toll gate. The toll gate system is considered to bear a very high infrastructure cost, and the users are the ones who help recover the cost through increments in their entrance bills.

ATG is said to use a combination of mobile telecommunication technology (GSM) with satellite-based Global Positioning System (GPS). Using GPS technology, the distance driven in kilometers can be estimated, and use to calculate the entrance fees and rates, and then transmit the information to the NATCS computer centre. Each vehicle is charged from the highway entrance up until the end of the highway. In order to identify the plate numbers of trucks, the system has control gates equipped with digital short range communication (DSRC) detection equipment and high resolution cameras [10,24]. The system is considered expensive due to the technical specifications which incur high cost for motorists.

Other systems include TouchNGo and SmartTAG[11]. This system uses IR technology, making it very vulnerable to failure. It also has high cost of implementation since users have to own two-piece tag required for this system. Passive RFID technology based toll gate system is another automated system that guarantees increased efficiency, since RFID is considered a highly stable technology with the elimination of human interaction in systems based on this technology [1].

1) Smartcard Based Toll Gate Automated System

It is observed that the use of contact type smart cards cannot be underestimated in the world of technology because it is being utilized for different purposes. The latest technology trends introduced contactless smartcards. They work on the *RF* frequencies. With the help of these smartcards there is no need to insert the smartcard in the reader, the reader reads the smartcard from the distance, and both the smartcard and the smartcard reader will transmit and receive signals which led to mutual transfer of information to other devices. It is considered faster than the contact based smart cards [2]. The Smartcard based toll gate automated system is considered effective and efficient since the card is recharged with some amount and whenever a person wants to pay the toll gate tax, just needs to insert the smart card and deduct amount using keypad, the system is security conscious since there is no need to carry cash. But it is considered expensive to install and maintain [2].

2) Rfid Based Toll Gate Automated System

Radio Frequency Identification (RFID) is an automated identification technology which uses Radio Frequencies between 30 kHz and 2.5GHz to identify objects remotely. The RFID Automatic toll gate system is designed to automatically detect the identities of the vehicles and perform the billing in accordance with the identity of each vehicle as pre-recorded in the database [12]. The system could automatically open and close and automatically emails the owners of the vehicles. These were the major achievements of the system. Other features are the ability to track vehicles and connect database remotely.

In spite of these, the system has failed in some of the required criteria because it did not yield the required result due to lack of resources and high cost of implementation. For instance, remote database connection needed a pre-set Virtual Private Network and automatic synchronizing software which will be readily available [12,28].

B. License Plate Detection

The development of a reliable and accurate License Plate Recognition (LPR) system cannot be underestimated in view of its potential application in traffic monitoring systems and highway entrance collection. LPR systems have recently attracted considerable interest as part of an Intelligent Transport System. While much commercial work has been done for Iran, Korean, Chinese, European and US license plates, little work has been done in LPR systems for developing country such as Nigeria [13]

The central idea of the license plate extraction is to detect vehicle plate number from video Existing systems are based on four modules. In the first module, the camera captures video of the vehicle. In the second module, the video is converted into frames by using MATLAB operations. The third module converts frames into images. Finally, in the last module by using canny edge detection and morphological operator's vehicle number plate is extracted. The main advantage of this technique is that from 10 sec video, 240 frames or images are extracted. Therefore, algorithms are implemented upon 240 images one by one automatically by the system. This serves as a major disadvantage because the system may need more computations [14].

Vehicle License plate identification is an essential stage in intelligent traffic system. In general, LPR comprises of four

stages: Image acquisition and processing, License plate extraction, License plate segmentation and License plate recognition [13]. The video image processing technology is commonly used to identify vehicles by their license plates. Real time LPR play a major role in automatic monitoring of traffic rules and maintaining law enforcement on public roads. There are different techniques such as Sobel edge detector [15,27], canny edge detection and morphological operators [16], skew correction [17] and color model [18]. One similarity between techniques discussed above is that they are implemented on single image and at day time only.

Canny edge detector [19,25] is an edge detection operator that uses a multi-stage algorithm to detect wide range of edges in images. Mathematical morphology is used as tools to extract image components that are useful in representation and description of the object shape [16, 26]. Analysis is based on set theory, topology, lattice algebra, function etc. An existing system referred to as the Vehicle License Plate Localization and Recognition System was developed based on digital images and could be easily applied to commercial car park systems for the use of documenting access of parking services, secure usage of parking houses and also to prevent car theft. The license plate localization algorithm was based on a combination of morphological processes with a modified Hough Transform approach and the recognition of the license plates was achieved by the implementation of the feed-forward back propagation artificial neural network [20,24].

Vicar is a neural network based artificial vision system able to analyse the image of a car, locate the registration plate and recognize the registration number of the car. The main features of the system are: it controls the stability-plasticity behaviour, it enhances reliability of the threshold, both off-line and online learning, self-assessment of the output reliability and high reliability based on high level multiple feedbacks. It has an OCR engine [21]. Other methods demonstrate the use of dynamic neuro-fuzzy model which enhances the prediction capability of the system and hence gives accurate estimation for adoption and selection of the new technology. Combining the strengths of neural networks and fuzzy logic, the process was relatively simple, supports creation of high level pedagogical strategies and could be easily adapted to individual technological preferences. Compared to neural networks, the neuro-fuzzy methods provide models which can be interpreted by human beings. The system is in the form of familiar if-then rules implying easy selection with the operators [22].

The major setback of the existing systems are categorized into cost, efficiency, accuracy, ease to use, applicability, availability of resources etc. Advanced video detection is capable of monitoring several lanes simultaneously from a side mount position and delivers an image of the traffic situation, but performance suffers from poor lightening and bad weather conditions. In addition, real-time video detection is computationally expensive and relies on high-performance signal processing hardware, large amounts of memory, and high-bandwidth data links. The density of installation along a highway route is limited and may not reach the density necessary to acquire enough data for reliable traffic flow or intensity prediction

III. Traffic Intensity and Estimation

Traffic estimation focuses on the population of vehicles speed, delays and queue lengths that result from the adoption of several traffic estimation strategies on individual roads, intersections and as well as toll gates. Intersection control and analysis tools throughout the world consider the population, speed, delays and queues as principal performance needed to be measured in order to determine the intersection level of service (LoS). This has to do with the evaluation of the adequacy of lane lengths, and the estimation of fuel consumption and emissions [23].

The rationale for concentrating on descriptive models is that a better understanding of the interaction between demand that is, arrival pattern and supply at traffic estimation is a prerequisite to the formulation of traffic control strategies. Estimation is based on assumptions regarding the characterization of the traffic arrival and service processes.

A. Design Considerations

The model is divided into three primary stages: data capture and pre-processing, data reduction and processing, and storage and feature analysis. In the data capturing stage, operations are performed to make data reduction and processing easier. The uncalibrated camera captures the in-coming vehicle's data as video stream and converts it to indexed images. The first processing step is done by converting those images to grayscale. Data reduction stage involves eliminating unnecessary generated text and duplicate plate numbers as shown in figure *1*. This stage includes finding the vehicle object in the frames, tracking it through the different frames and calculating the speed of the vehicles. All necessary information such as the arriving speed value, date, time, are also recorded in the database.

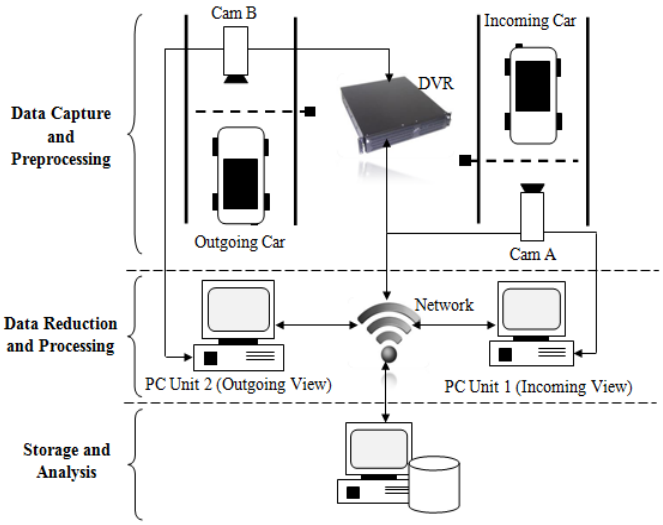


Fig. 1. A Strategic Traffic Monitoring System (STMS)

Capturing of video is done with the aid of the uncelebrated cameras. The cameras were strategically setup to capture the license plate as each car move across the camera lens. Closed Circuit TV (CCTV) was used for data capturing. The extraction of license is done with an Optical Character Reader (OCR) engine which is pre-installed on the system. The engine

is responsible for capturing the available vehicle from a suitable frame of the video file.

B. Traffic Monitoring System

The Traffic Monitoring System is an encapsulation of basically two modules namely; Traffic information generation (TIG) module and the information analysis (IA) module. TIG is responsible for generating all the necessary information that will be needed to estimate the traffic during the analysis stage. It involve getting the video either in real time or pre-recorded from the Digital Video Recorder (DVR), detecting and extracting license plate number and storing up alongside the arrival speed, date and time in the database. Storage in the Traffic Monitoring System is categorized into two, storage of generated license plate number and other information about each vehicle after reading the video signal at real time. The instructions for TIG are given in algo *1*.

Algo 1: Algorithm for TIG
Step 1: *Get video file from storage*
Step 2: *Convert video to still frames*
Step3: *Detect and Identify frame with valid license plate number*
Step 4: *Get speed value of the host vehicle*
Step 5: *Extract license plate number from frame*
Step 6: *Verify license plate number existence in database*
Step 7: *If (verification result = true)*
(Return "request check IN or OUT") Else
(Save Vehicle record (license_plate_number, speed value, date, time, cap_Time in database)
Step 8: *saved license_plate_number with speed value, date and time and capture time.*

Traffic Information Analysis (TIA) module makes reference to the database and gets all necessary information needed to generate traffic report about a selected data provided it exist in the database. To generate the report, information about the speed, total number of license plate, total time etc. are needed. Algo *2* describes the steps involved to illustrate the Traffic Information Analysis Module:

Algo 2: Algorithm for TIA
Step 1: *Get date*
Step 2: *Verify the existence of record for the selected date*
Step 3: *If (Verification = false)*
(Return "No Record found") Else
(Proceed to next step)
Step 4: *Total Speed (Ɛ) = Total Speed + Speed Value 1 + Speed Value 2 + ... + speed n*
Step 5: *Average speed = Total Speed (Ɛ)/ Total number of existing License_plate_number (n)*
Step 6: *Total_time (T) = Total time + cap_Time 1 + cap_Time 2 + ... + cap_Time*
Step 7: *Traffic_Intensity Value (β) = Total number of existing license plate (n) / Total_time (T)*
Step 8: *Return Remark ("Minimal Traffic")*
If (n <= 50 per minute)
Else if (n > 50 per minute)
Return Remark ("Moderate Traffic")
Else if (n > 100 per minute)
Return Remark ("Intense Traffic")
Step 9: *Generate Traffic report*
Step 10: *Associate generated report with bar chart.*

The traffic monitoring system (TMS)flow structure shows the basic implementation of the algorithms in the license plate number and report generation presented in figure *2* and the use-case model for the TMS is described in figure *3*. Figure *3* illustrates the overall accessibility of the different personnel and users of the system resources. It also illustrates the access control features of three major personnel namely: Administrator, Traffic Information Officer and the Security Officer. The administrator is considered to have no limits in accessing all the system resources. The administrator is responsible for the overall efficient and functional running of the system. He controls the activity of the system from both the front and back end. He gives report to the Traffic Information Officer (TIO).

Based on hierarchy, the Traffic Information Officer is considered the second level of access to the system resources. The access of the TIO is limited to viewing traffic report and information in the Database. The Security Officer has the lowest accessibility hierarchy to the security office in the front view.

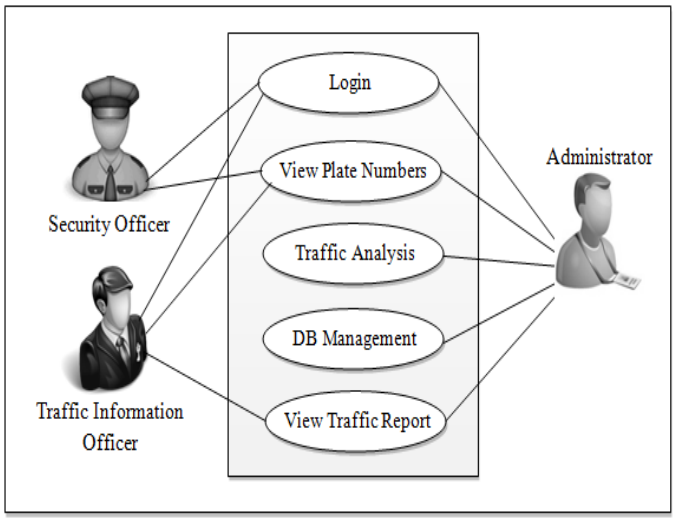


Fig. 2. License Plate number and Traffic information generation

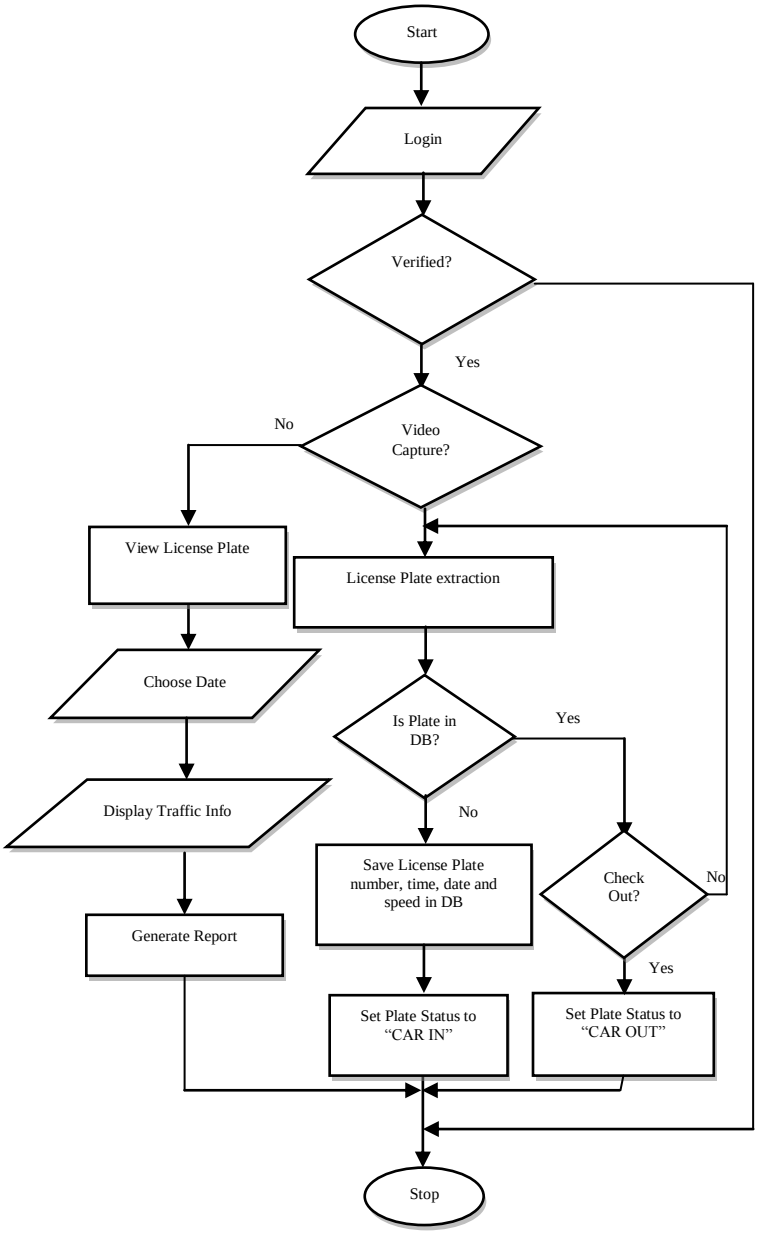


Fig. 3. Use case access control for Traffic Monitoring

C. Toll Gates Traffic Intensity in Queueing Networks

Traffic Intensity is a measure of the average occupancy of a server or resource during a specified period of time, normally a busy hour. It is measured in traffic units (erlangs) and defined as the ratio of the time during which a facility is cumulatively occupied to the time this facility is available for occupancy. If a toll gate is viewed as an interaction of services which vehicles pass through sequentially, then it is natural to model the system as a queueing network.

A queueing network can be represented as a directed graph shown in figure *4* in which the nodes denote the service points. Each point of the gate is represented by a separated service facility called a service centre. The ability of the service centre to provide services for any arriving vehicles does not depend only on the mean arrival rate but also on the pattern in which they arrive. The service time is the time a vehicle spends at the toll gate before proceeding. If the average duration of a service interaction between a service point and a vehicle is $\frac{1}{\mu}$, then μ is the service rate of the toll gate.

If there are more than one service points as observed in most toll gates the processing of vehicles can run parallel at the same time: one vehicle at each service point server. If there are fixed service points, say in this case *4*, then it is called multiple service points of servers *c=4*, each of which can service a vehicle at any time.

Figure *5* shows a multiple queue model in which vehicles at any point(s) which cannot receive service immediately, automatically waits in a buffer. The arrival rate λ and the service rate are the most important features of a single queue. And the traffic intensity ρ is denoted as

$$\rho = \frac{\lambda}{cx\mu} \quad (1)$$

A situation where the arrival points are different from the departures is considered. A transition diagram is described in figure *6*.

If the arrival rate at queue *i* is λ_i so also the departures will experience a Poisson departure stream with queue *i* is *i+1* and λ_{i+1} also becomes λ_i at another arrival. By decomposition principle, when the departure stream is split in this way, each of the resulting arrival streams will also be Poisson. If the service centres of any queue are analysed in this respect, expressing its input as a sum of output streams, what is obtained is called traffic equations. The traffic equation for figure *5* is as follows:

$$\lambda_1 = \lambda + q + \lambda_2$$

$$\lambda_2 = px\lambda_1 \quad (2)$$

One equation for each service point will give one unknown and this will end up with *n* equations with *n* unknowns. This traffic equation can be solved to get the arrival rate at each service point. This led to the utilization of the queue. The queue is being utilized whenever it is non-empty. Therefore, the Utilization *U*, is $1-\pi_0$ which implies $U = \rho$.

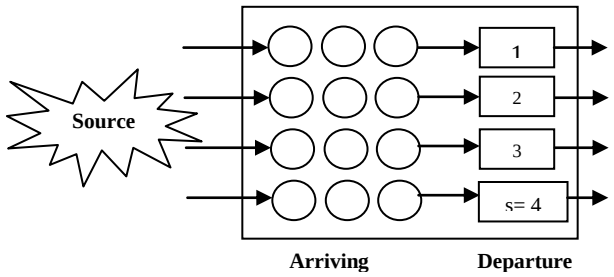


Fig. 4. An Toll gate Open Queueing Network

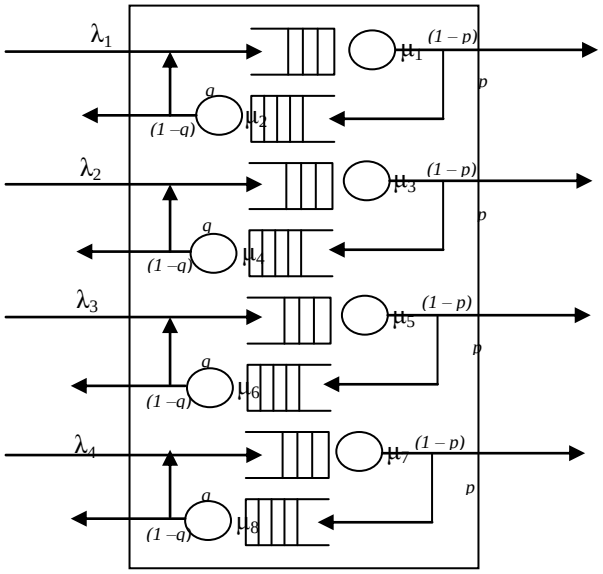


Fig. 5. Multiple Queues at Toll gate

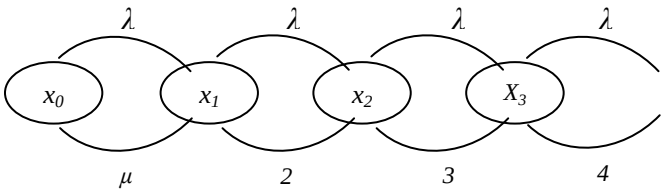


Fig. 6. An Toll Gate State Transaction Diagram for M/M/c

IV. Implementation and Results

The integral subsystem of TMS consists of four main parts namely; Cameras, Computer (PC), Network Devices, Database. The cameras are uncalibrated video cameras that take in video signal as input data for the system. The camera reads in the video for processing in real-time and record overtime for batch processing. The PC hosts the software for processing and the network devices create connectivity between PCs indicating incoming and outgoing views. The unprocessed and processed information are stored by the database. The Object Oriented Paradigm (OOP) designed with the front end implemented with C# with Microsoft .NET considering it flexibility, provides an interface for viewing the activities of vehicle at the entrance and it is capable of analysis after taking in all necessary data correctly. It also provides an easy approach for debugging and correction of errors. The backend is designed using the Microsoft SQL Server connected to the C# modules.

The recorded is stored in the DVR and conversion of the video frame to an image was performed by the OCR engine. This process involves converting the single movie frame into an indexed image and an associated color map. A simple loop is used to repeat this operation for the entire movie, the dimension of the loop is decided by the length of the video and in turn return the total number of frames in the movie.

The Image Index is changed to its grays clay equivalent in order to prepare the index image for better identification of characters in the license place received. Bounding box extraction is followed by calling the property of the OCR engine to isolate the area of interest in the license plate image. The extraction of license is done with an OCR engine which is pre-installed on the system to be used. The engine is responsible for capturing the available license plate character from the suitable frame of the video file.

A. Application Description

The Administrator has complete control over the resources of the system. The access control is made possible with different accessing passkeys. The administrator's key enables all features of the application, but other keys only activates some features and disable others depending on who is logging in. Figure *7* consists of video feedback that shows camera recording in real time. The Information are: license per time, the capture time, the estimated arriving speed as well as the departure time of the vehicles. Figure *7* shows buttons used to access the list of captured license plate numbers. The buttons are enabled depending on who is logged in to access some control features. A multi-stage algorithm in canny edge detector was used to detect the vehicle images; result is shown in figure *8*.

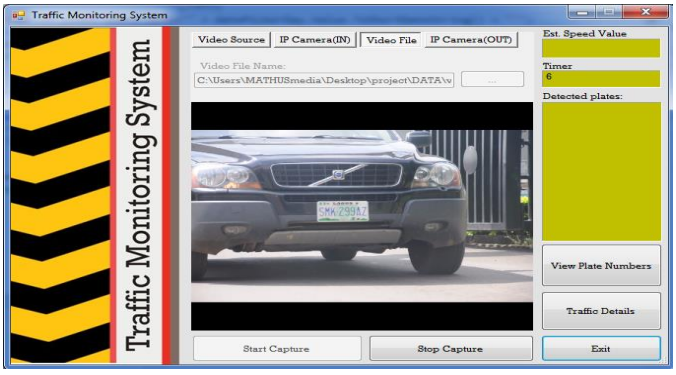


Fig. 7. Video Feedback of incoming Vehicles

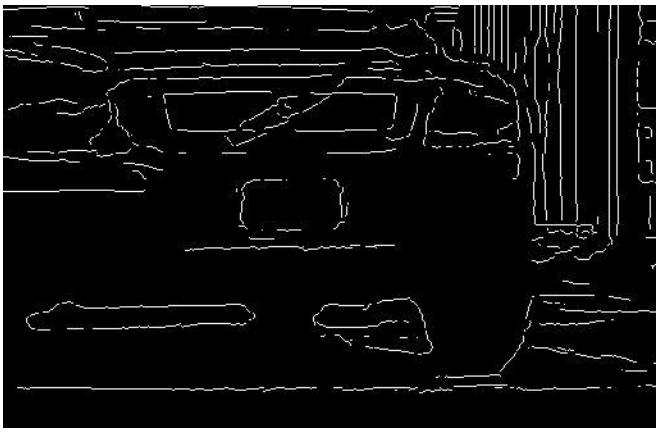


Fig. 8. Canny Edge Detection of the Vehicle on queue

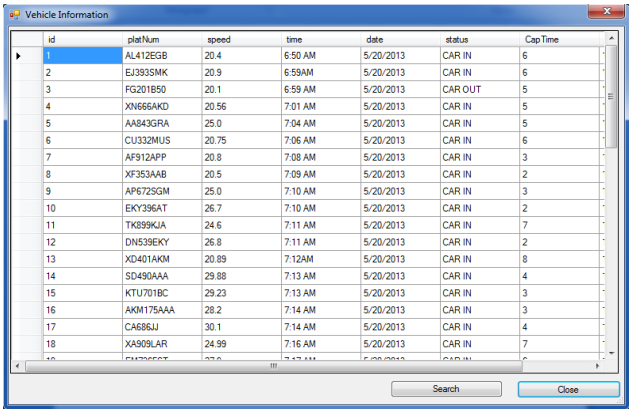


Fig. 9. Incoming Vehicles Information

The vehicle information is displayed in figure *9*. It is accessed with view plate's button from the main interface. It provides information about the arriving speed, time, date and time of arriving and departing ones.

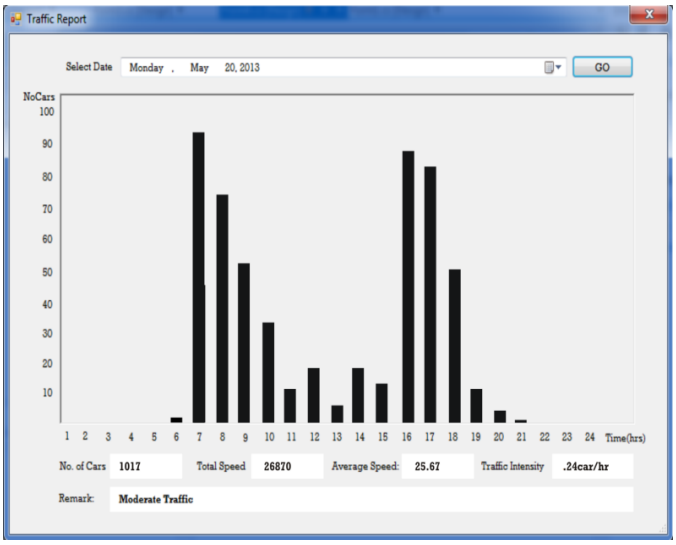


Fig. 10. Generated Report for Traffic Estimation

Figure *10* shows the traffic reports accessed with the traffic details button. The user selects the required date from the date selector box to generate the traffic report for the selected date. The histogram in figure *10* generated results for the number of cars per hour and the traffic information values for the total number of cars, total speed, average speed and the value for traffic intensity per day.

B. Discussion

The results were generated based on two days investigation carried out at the Federal University of Agriculture, Abeokuta, Nigeria (FUNAAB) ceremonial gate. The ceremonial gate officially open by 6:00am and closes by 10:00pm. The data were captured by pre-recording the traffic activity of the ceremonial gate with permission from the University Chief Security Officer (CSO). An uncalibrated mini DV video camera is used to capture the vehicle activities and giving an AVI format files, which are then batch processed by the software application. The result generated from the observation is given on tables *1* and *2*.

TABLE I. CAR PER HOURS FOR DAY 1

Time Range	No. of Incoming Vehicles	No. of Outgoing Vehicles
6:00am – 6:59am	21	0
7:00am – 7:59am	142	6
8:00am – 8:59am	122	5
9:00am – 9:59am	81	10
10:00am – 10:59am	29	5
11:00am – 11:59am	24	10
12:00pm – 12:59pm	8	22
1:00pm – 1:59pm	12	25
2:00pm – 2:59pm	14	53
3:00pm – 3:59pm	12	60
4:00pm – 4:59pm	20	169
5:00pm – 5:59pm	9	83
6:00pm – 6:59pm	8	35
7:00pm – 7:59pm	3	20
8:00pm – 8:59pm	-	-
9:00pm – 9:59pm	-	-

Figure *9* shows the traffic information at the FUNAAB ceremonial gate on 20th May 2013. It illustrates the trend at which the traffic varies with time. From the figure it could be observed that there is more vehicles coming between 7:00am and 9:00am and the number decreases as it move towards midday. At the outgoing end, it observed that there is gradual increase in the number of vehicles going out from midday towards evening. The same set up used on day 1 was repeated on day two.

Figure *10* illustrates the traffic information on day two. From the figure, it is observed that the trend was similar to the day one showing more cars arriving between 7:00am to 9:00am and most cars leaving between 3:00pm to 6:00pm. But there exist a significant difference, it could be observed that on this particular day more cars came in early compared to day one which is around 6:00am to 7:00am.

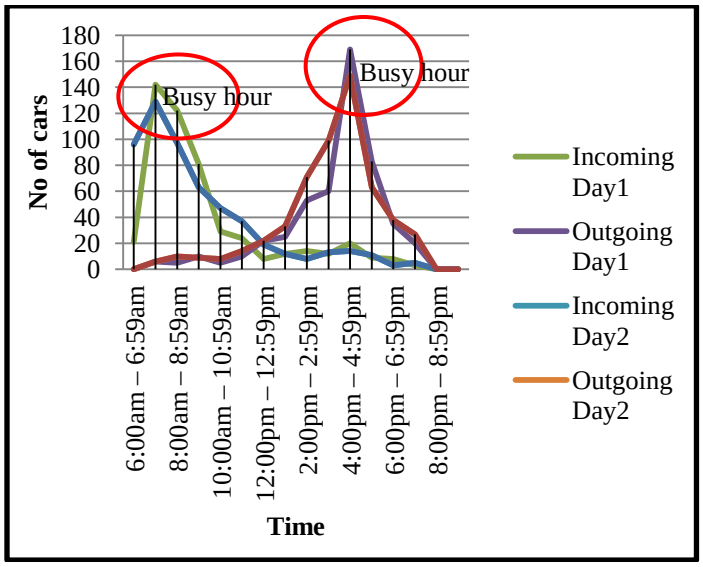


Fig. 11. Traffic Variations Showing the Busy Hours

Figure *11* illustrates the overall activities at the ceremonial gate for the two days of observation. It could be observed that the peak of the graph is at 7:00 am – 9:00am on the incoming side and 3:00pm – 5:00pm respectively. Hence it is assumed that the busy hours of the ceremonial gate are between 7:00-9:00am and 3:00-5:00pm respectively.

C. Evaluation

The evaluation is done using average percentage of success and the processing time. The throughput of the system was also examined to further examine the rate of achievement in determining level of success.

1) Average Percentage of Success and Processing Time

To examine the efficiency of the system, an average percentage of success tests with respect to average processing time is used. Average percentage of success is the percentage value of the total number of license plate generated by the application divided by the total number of vehicles counted manually ensuring 100% accuracy. To calculate the average percentage of success the formula is given as:

$$APS\ (\%) = \frac{NLG}{NMC} \times 100$$

APS (%) = Average Percentage of Success
NLG = Total number of license of license plate generated by application software
NMC = Total number of vehicle generated manually

The result displaced in table 2 was derived using the above formula choosing sample time of 7:00am – 10:00am and 3:00pm – 6:00pm. The results are shown in table 5.

TABLE II. AVERAGE PERCENTAGE OF SUCCESS AND PROCESSING TIME

Sample Hrs.	DAY1 Incoming (%)	DAY1 Outgoing (%)	DAY1 Cap. Time (secs)	DAY2 Incoming (%)	DAY2 Outgoing (%)	DAY2 Cap. Time (secs)
7:00am - 7:59am	99.9	100.0	4.1	99.2	100.0	4.3
8:00am - 8:59am	93.1	99.7	3.2	83.1	99.1	5.0
9:00am - 9:59am	86.1	100.0	5.0	100.0	99.0	4.3
3:00am - 3:59am	98.6	99.7	3.7	100.0	97.8	2.8
4:00am - 4:59am	100.0	89.3	2.1	99.5	99.3	3.1
5:00am - 5:59am	99.5	96.7	3.6	100.0	99.7	4.3

2) Throughput

Figure 12 shows that the average percentage of success of the application capturing the saving license plate falls between 80% and 100%. The failures are only due to factors such as faded license plate, low light etc. Figure 12 also shows that the maximum capture time of *5* seconds validate the speed and processing time of the application which is considered very fast under the required hardware specifications. The throughput for this process is drawn in figure 12. The result shows that at the busy hours; between 6am - 9am and 4pm-6pm, the throughput is very high otherwise very low throughput is experienced.

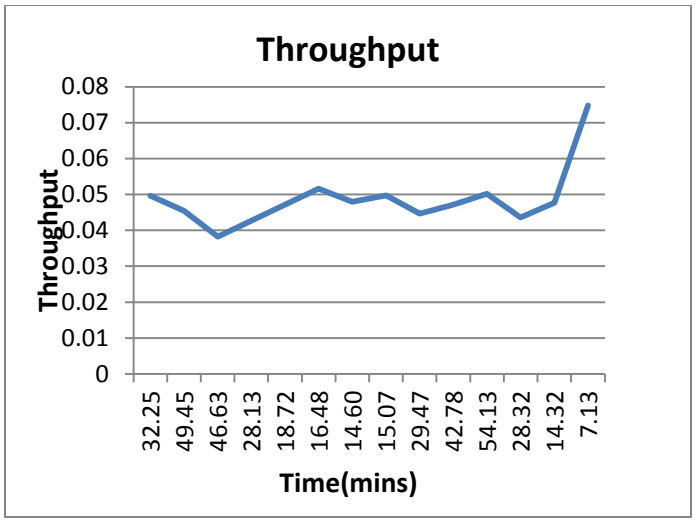


Fig. 12. Throughput versus Time

V. CONCLUSION

This research work presents a simple, low cost and efficient traffic monitoring system that allow the automated monitoring of traffic activity at a toll gate in which the system is implemented. The operation of TMS is based on the knowledge of the license plate number generated by the software application. The software makes reference to the number of generated license plate number with respect to time taken to generate all the number and in turn estimate the traffic intensity. This system is considered very efficient in detecting vehicle number plate from video files, either in real time or pre-recorded, all with the aid of the OCR engine in the application. It is highly effective in security areas like border crossing, entrance plaza, school gates, petrol station etc. It shows number plate automatically extracted from all images.

The estimated accuracy of capturing vehicle number plate from number of frames/images in a single video is 85% to 90%; while getting desired result of number plate from overall video is 99%. The advantage of this work is optimal in the provision of a simple automated, low cost, efficient and effective system of monitoring traffic and the idea of how license plate number can be used in the estimation of traffic intensity at toll gates. However, major problems posed are the inability of the application software to read the character in faded or complete washed off license plate.

Monitoring traffic is considered necessary in order to keep record and make traffic/vehicle related decisions. The limitations of the manual record keeping with the aid of the tally and the high cost of implementing the existing automated system gives room for the development of a simple and low-cost automated system to perform this tasks with more efficiency. TMS allows an automated license plate extractionto take place and in turn keeping other records about the vehicle hosting the license plate simultaneously. This reduces the manual labour and delays that often occur in the using of the manual tally system. This system of collecting entrances is eco-friendly and also results in increased entrance lane capacity.

Moreover, the system can generate report automatically, which makes it is convenient for the user. The proposed system does not necessarily require a professional operator to work, it has a simple GUI. The Traffic Monitoring systemis efficiently used in the traffic surveillance for detection since it is capable of identifying individual vehicle with their license.

REFERENCES

[1] Khadijah, K. and Ismail, W., "Electronic Entrance Collection System Using Passive RFID Technology" Journal of Theoretical and Applied Information Technology, 2005, 70-76.

[2] Raadhikaa, L."The Universal Journal of applied Computer Science and Technology, UNIASCIT, 1, 1, 2011, 05-08.

[3] Friedman, J., " Finding Mean Traffic Speed in Low Frame-RateVideo, IEEE Trans. Intelligent Transportation Systems 1, 2, 2005, 98-107.

[4] Lazaros G., George K., Elli P., "Automatic Estimation of Vehicle Speed from Uncalibrated Video Sequences", International symposium on modern technologies, education and professional practice ingeeodesy and related fields,1-5, 2005.

[5] Chetan, S. and Amandeep, K., " Indian vehicle license plate extraction and segmentation", InternationalJournal of Computer Science and Communication, Dept. of CSE, Patiala, 2, 2, 2011, 593-599.

[6] Lauren S. and Mariko B.,"Electronic Entrance Collection 2007.

[7] Xu, L., Wei, S. Ming-Lu, L. and Min-You, W., "Vehicle-based Sensor Networks for Traffic Monitoring", 2010.

[8] Xie, H., Kulik, L. and Tanin., E.,"Privacy Aware Traffic Monitoring", 2011, 1-7.

[9] Khali, P., Michael, C.W. and Shahriyar, H.,"Entrance Collection Technology and Best Practices", Project 0-5217:Vehicle/License PlateIdentification for Entrance Collection Application, 2007.

[10] Gabriel, N., Mitraszewska, I. and Tomasz, K, "The Polish Pilot Project of Automatic Entrance Collection System", Proceedings of the 6thInternational Scientific Conference TRANSBALTICA, 2009.

[11] Ismail, M.S., KhairUl-Anwar, M.Y. and Zaida, A.Z., "Electronic Entrance Collection (ETC) Systems Development in Malaysia", PIAR CInternational Seminar on Intelligent Transport System (ITS) in Road Network Operations, 2006.

[12] Gunda, L., Lee M., Reginald G., Mhlanga, S. and Nyanga L., "Vehicle License Plate ", CIE42 Proceedings, Cape Town, South Africa 2012.

[13] Jameel, A.,"License Plate Recognition System, Thesis Presented to the Deanship of Graduate Studies", King Fahd University of Petroleum and Minerals, 2003, 2-5.

[14] Kaur, H., Manviand Balwinder, S., "Vehicle License Plate Detection from Video Using Edge Detection and Morphological Operators". Singh International Journal of Engineering Research & Technology ISSN: 2278-0181, 1, 9, 2012.

[15] Suri, P. K., Ekta, W. and Amit, V., "Vehicle Number Plate Detection using Sobel Edge Detection Technique", UCST, 1, 2, 2010, 178-182.

[16] Vasudevan, Shriram, K., Dharmendra, T. and Sivaraman, R., "Automotive Image Processing Technique using Canny's Edge Detector" International Journal of Environmental Science and Technology, 2010, 2632-2644.

[17] Manchikalapudi, V., "Skew Correction and Localisation of Number Plate Using Hough Transform" UCST, 2011, 472-476.

[18] Deb, K. and Hyun Jo, K. "Segmenting the License Plate Region using Color Model", University of UlsanSouth Korea, 2009, 401-418.

[19] Little, J. D. C.. "A Proof for the Queuing Formula: L = λW". Operations Research**9,** 3,1961, 383–387.

[20] Ganapathy, V. and Lui, D. A., "Malaysian Vehicle License Plate Localization and Recognition System", School of Engineering, Monash University of Malaysia, 2007.

[21] Draghici, S., " A Neural Network-Based Artificial Vision System for License Plate Recognition",International Journal of Neural System. 8, 1, 1997, 113–126.

[22] Kalbande, D.R., Signal, P., Denotable, N., Shah, S. and Tampa, G.T.,"An Advanced Technology Selection Model using Neuro Fuzzy Algorithm for Electronic Entrance Collection System", International Journal of Advanced Computer Science and Applications, 2, 4, 2011.

[23] Rouphail N., Tarko, A. and Li, J., "Traffic Flow at Signalized Intersections" Transportation Research Record, 2009, 9-32.

[24] Dailey, D.J. And Li, L "An Algorithm to Estimate Vehicle Speed using Uncalibrated Camera", IEEE, vol 4,8, 3-12. 1999.

[25] Leung, T. and Malik, J., "Representing and recognizing thevisual appearance of materials using three-dimensional textons. IJCV, 43, 1, 2001, 29-44.

[26] Lowe, D., " Distinctive image features from scale-invariant keypoints". IJCV 2004.

[27] Lowe, D. G., " Object recognition from local scale-invariant features. InICCV, 1999,1150-1157.

[28] Paulev_e, L., J_egou, H., and Amsaleg, L., " Locality sensitive hashing:a comparison of hash function types and querying mechanisms. Pattern Recognition Letters,31,11, 2010, 1348-1358.

Performance Analysis of Faults Detection in Wind Turbine Generator Based on High-Resolution Frequency Estimation Methods

CHAKKOR SAAD
Department of Physics, Team: Communication and Detection Systems, University of Abdelmalek Essaâdi, Faculty of Sciences, Tetouan, Morocco

BAGHOURI MOSTAFA
Department of Physics, Team: Communication and Detection Systems, University of Abdelmalek Essaâdi, Faculty of Sciences, Tetouan, Morocco

HAJRAOUI ABDERRAHMANE
Department of Physics, Team: Communication and Detection Systems, University of Abdelmalek Essaâdi, Faculty of Sciences, Tetouan, Morocco

***Abstract*—Electrical energy production based on wind power has become the most popular renewable resources in the recent years because it gets reliable clean energy with minimum cost. The major challenge for wind turbines is the electrical and the mechanical failures which can occur at any time causing prospective breakdowns and damages and therefore it leads to machine downtimes and to energy production loss. To circumvent this problem, several tools and techniques have been developed and used to enhance fault detection and diagnosis to be found in the stator current signature for wind turbines generators. Among these methods, parametric or super-resolution frequency estimation methods, which provides typical spectrum estimation, can be useful for this purpose. Facing on the plurality of these algorithms, a comparative performance analysis is made to evaluate robustness based on different metrics: accuracy, dispersion, computation cost, perturbations and faults severity. Finally, simulation results in MATLAB with most occurring faults indicate that ESPRIT and R-MUSIC algorithms have high capability of correctly identifying the frequencies of fault characteristic components, a performance ranking had been carried out to demonstrate the efficiency of the studied methods in faults detecting.**

Keywords*—Wind turbine Generator; Fault diagnosis; Frequency Estimation; Monitoring; Maintenance; High Resolution Methods; Current Signature Analysis

I. INTRODUCTION

The increasing demand in energy over the world, as well as the growth in the prices of the energy fossil fuels resources and it exhaustion reserves in the long run, furthermore the commitment of the governments to reduce greenhouse gases emissions have favored the research of others energy sources. In this context, the recourse to renewable energy becomes a societal choice. The development of this alternative is encouraged because it offers natural, economic, clean and safe resources. Among the renewable energies, wind energy which has been progressed in a remarkable way in these recent years. It provides a considerable electrical energy production with fewer expenses with exception of construction and maintenance budget. Actually, wind energy investment has increased by multiplication of the wind parks capacities.

This contributes greatly to the expansion of terrestrial and offshore wind parks. These parks are usually installed in far locations, difficult to access, subject to extreme environmental conditions. Therefore, a predictive monitoring scheme of wind turbines, allowing an early detection of electromechanical faults, becomes essential to reduce maintenance costs and ensure continuity of production. It means that stopping a wind installation for unexpected failures could lead to expensive repair and to lost production. This operating stopping becomes critical and causes very significant losses. For these reasons, there is an increase need to implement a robust efficient maintenance strategy to ensure uninterrupted power in the modern wind systems preventing major component failures, facilitating a proactive response, minimizing downtime and maximizing productivity [1], [8]. To anticipate the final shutdown of wind generators, on-line condition monitoring would be the most efficient technique because it allows the assessment of the health status of an operating machine by analysis of measured signals continuously [8]. Different types of sensors can be used to measure physical signals to detect the faults with various existing methods [4], [7], [9], [10], [12].

This is why reliability of wind turbines becomes an important topic in scientific research and in industry.

Most of the recent researches have been oriented toward electrical monitoring, with focus on the generator stator current. One of the most popular methods for fault diagnosis is the current signature analysis (CSA) as it is more practical and less costly [3], [4], [9], [10], [12]. Within the last decade many studies based on signal processing techniques have been conducted to detect electric machine faults prior to possible catastrophic failure. These researches initially developed for electric motor can be easily adapted to wind turbine generator. Furthermore, with recent digital signal processor (DSP) technology developments, motor and generator fault diagnosis can now be done in real-time [3]. Among signal processing techniques, non-parametric, parametric and high resolution or subspace methods (HRM) are widely adopted in machine diagnosis. They can be used for spectral estimation [10], [15], [16], [23], [24]. However,

This research work carried out in this direction with subspaces methods not highlight metrics of accuracy, robustness level of each approach related to the failures severity and computation time which is a key parameter in the context of a real-time integration. Otherwise, an investigation focused on the mean square error (MSE) and on the variance of faults harmonic detection must be done to evaluate the accuracy and detection robustness especially when the parameters of the signal, containing the faults in formations, will changes according to constraints of the application [18]. The main object of this study is to search a robust high resolution detection method for condition supervision, suitably adapted for implementation in wind generator.

II. Related Work

In the literature review, many research studies applying enhanced signal processing techniques and advanced tools have been commonly used in the wind generator stator current to monitor and to diagnose prospective mechanical or electrical faults. As known, these faults cause a modulation impact in the magnetic field of the wind generator, which is reflected by the appearance of a significant harmonics (peaks) in the stator current spectrum [8]. Nevertheless, these techniques are inappropriate because they have drawbacks such as high complexity, poor resolution and/or may suffer from some limitations. However, some failures are characterized by non-stationary behaviors [8], [14]. For this reason some researchers are leaning particularly toward methods adapted for non-stationary signals, such as time-frequency analysis, spectrogram, the wavelet decomposition (scalogram), Wigner-Ville representation, Concordia Transform (CT) and the Hilbert-Huang transform [13], [31].

In the first hand, in [7] a statistical diagnosis approach is proposed based on residues analysis of the electrical machine state variables by the use of the Principal Components Analysis method (PCA) for faults detection in Offshore Wind Turbine Generator. The aim drawback of this approach is that the detection efficiency requires a good choice of the principal components number. Some researchers are proposed failures diagnosis of wind turbines generators using impedance spectroscopy (IS) [27]. On the other hand, the periodogram and its extensions which are evaluated through a Fast Fourier Transform (FFT) is not a consistent estimator of the PSD because its variance does not tend to zero as the data length tends to infinity. Despite of this drawback, the periodogram has been used extensively for failure detection in research works [12], [17]. The (FFT) does not give any information on the time at which a frequency component occurs. Therefore, the Short Time Fourier Transform approach (STFT) is used to remove this shortcoming. A disadvantage of this approach is the increased sampling time for a good frequency resolution [32]. The discrimination of the frequency components contained within the signal, is limited by the length of the window relative to the duration of the signal [25]. To overcome this problem, in [8] and in [13] Discrete Wavelet Transform (DWT) is used to diagnose failures under transient conditions for wind energy conversion systems by analyzing frequencies with different resolutions. This method facilitates signal interpretation because it operates with all information contained in the signal by time-frequency redistribution.

One limitation of this technique that its gives a good time resolution and poor frequency resolution at high frequencies, and it provides a good frequency resolution and poor time resolution at low frequencies [12], [28]. Due mainly to their advantages, in [26] parametric methods have improved performance though they are affected by an adequate signal to noise ratio (SNR) level. High resolution methods (HRM) can detect frequencies with low SNR. They have been recently introduced in the area of induction motors and wind generators faults diagnosis by the application of multiple signal classification (MUSIC) method [2], [6], [29]. MUSIC and its zooming methods are conjugated to improve the detection by identifying a large number of frequencies in a given bandwidth [2], [28], [30].

Moreover, eigen analysis methods are especially suitable in case that the signal components are sinusoids corrupted by additive white noise. These algorithms are based on an eigen decomposition of the correlation matrix of the noise corrupted signal. Another approach used is ESPRIT [20], [26] [33], [34], [35]. It allows and performs well determination of the harmonic parameters components with high accuracy. In fact, this paper investigates the most efficient high-resolution techniques to detect faults in wind turbine generator.

III. Faults In Wind Turbine Generator

The wind generator is subjected to various electro-mechanical failures that affect mainly five components: the stator, the rotor, the bearings, gearbox and/or air gap (eccentricity) [5]. These faults require a predictive detection to avoid any side effect causing a breakdown or a fatal damage. However, a recent literature surveys [36], [37] shows that these defaults require periodic monitoring to avoid any unforeseen deterioration. Recent researches have been directed toward stator current supervision. Particularly, the current spectrum is analyzed to extract the frequency components introduced by the fault. A summary of wind turbines faults and theirs related frequencies are presented in Table I.

TABLE I. Wind Turbines Faults Signatures

Failure	Harmonic Frequencies	Parameters
Broken rotor bars	$f_{brb} = f_0\left[k\left(\frac{1-s}{p}\right)\pm s\right]$	$k = 1,3,5,...$
Bearing damage	$f_{bng} = \left\| f_0 \pm k\, f_{i,o} \right\|$	$k = 1,3,5,...$ $f_{i,o} = \begin{cases} 0.4\, n_b\, f_r \\ 0.6\, n_b\, f_r \end{cases}$
Misalignment	$f_{mis} = \left\| f_0 \pm k\, f_r \right\|$	$k = 1,3,5,...$
Air gap eccentricity	$f_{ecc} = f_0\left[1 \pm m\left(\frac{1-s}{p}\right)\right]$	$m = 1,2,3,...$

Where f_0 is the electrical supply frequency, s is the per-unit slip, p is the number of poles, f_r is the rotor frequency, n_b is the bearing balls number, $f_{i,o}$ is the inner and the outer frequencies depending on the bearing characteristics, and m, $k \in \mathbb{N}$ [8], [12], [26].

IV. Wind Generator Stator Current Model

To study the mentioned faults detection methods, the current will be denoted by the discrete signal x[n], which is obtained by sampling the continuous time current every $T_s=1/F_s$ seconds. The induction machine stator current x[n] in presence of mechanical and/or electrical faults can be expressed as follows [26]:

$$x[n]=\sum_{k=-L}^{L} a_k \cos\left(2\pi f_k\left(\omega(n)\right)\times\left(\frac{n}{F_s}\right)+\varphi_k\right)+b[n] \quad (1)$$

Where x[n] corresponds to the n^{th} stator current sample, b[n] is a gaussian noise with zero mean and a variance equals to $\sigma^2= 10^{-4}$ i.e. $b[n] \sim (0, 10^{-4})$. L is the number of sidebands introduced by the fault. The parameters $f_k(\omega)$, a_k, φ_k correspond to the frequency, the amplitude and the phase of the k^{th} component, respectively. ω(n) is a set of parameters to be estimated at each time n depending on the studied fault. The time and space of harmonics are not considered in this paper. The problem to solve is treated as a statistical estimation problem. It is an estimation of the fundamental frequency, the characteristic faults frequencies, and their amplitudes by the computation of the current spectrum from the stator current samples x(n).

V. High-Resolution Frequency Estimation Methods

In this section, a brief description of each studied high resolution method and its main features are presented. The subspace frequency estimation methods rely on the property that the noise subspace eigenvectors of a Toeplitz autocorrelation matrix are orthogonal to the eigenvectors spanning the signal space. The model of the signal in this case is a sum of random sinusoids in the background of noise of a known covariance function. Among these methods, Prony method which is used for modeling sampled data as a linear combination of exponential functions. Although, it allows extracting P sinusoid or exponential signals from time data series, by solving a set of linear equations [15], [16], [24]. The signal s(n) is assumed equal to a sum of damped sines verifies the following recursive equation :

$$s(n)+b_1 s(n-1)+\cdots+b_{2P}s(n-2P)=0 \quad (2)$$

$$B(z)=z^{2P}+b_1 z^{2P-1}+\cdots+b_{2P} \quad (3)$$

Polynomial (3) has 2P complex conjugate roots given by:

$$z_k=\rho_k e^{\pm 2\pi j f_k} \quad (4)$$

It's possible to calculate b_k, then the roots z_k and therefore the frequencies f_k and the damping coefficients ρ_k.

Unlike the methods using the periodogram, even with windowing, the high resolution methods are such that the error tends to zero when SNR→ ♁

The Pisarenko Harmonic Decomposition PHD relies on eigendecomposition of correlation matrix which is decomposed into signal and noise subspaces. This method is the base of advanced frequency estimation methods. It has a limited practical use due to its sensitivity to noise [24], [15], [19], [22], [24], [33]. The eigenvector *v* associated with the smallest eigenvalue of the (2P+1) order covariance matrix R_x of the observation has, as its components, the coefficients of the recursive equation (2) associated with the frequencies of the signal s(n). Then, the 2P degree polynomial B(z) is constructed based on *v* [19], [39]. The 2P complex conjugate roots z_k are extracted from it, which leads us to the frequencies:

$$f_k=\frac{1}{2\pi}\arg(z_k) \quad (5)$$

For MUSIC (Multiple Signal Classification) approach, it is the improved version of Pisarenko method where M-dimensional space is split into signal and noise subspaces using many noise eigenfilters. The size of time window is taken to be M > P+1. Therefore, the dimension of noise subspace is greater than one and is equal to M-P. Averaging over noise subspace gives improved frequency estimation. Once the eigendecomposition of correlation matrix is calculated, it's used to find the (M x (M-P)) matrix G constructed from the (M -P) eigenvectors associated with the (M-P) smallest eigenvalues. Afterwards, the (M x M) matrix GG^H is calculated to find the coefficients of the polynomial equation [6], [21], [22], [23], [24], [29], [33], [38]:

$$\tilde{Q}(z)=\left[z^{M-1}\cdots\ z\ \ 1\right]GG\left[1\ \ z\ \cdots\ z^{M-1}\right]^T \quad (6)$$

Then, the estimation of the P frequencies values can be achieved as following:

$$f_k=\frac{\theta_k}{2\pi} \quad (7)$$

Two possibilities are available:

1) Calculating the 2(M-1) roots of $\tilde{Q}(z)$, then keeping the P stable roots that are closest to the unit circle. This is called the Root-MUSIC method.

2) Finding the P minima of $\tilde{Q}\left(e^{j\theta_k}\right)$, using FFT function. This is called the FFT-MUSIC method.

Another method is Eigenvector (EV), this technique estimates the exponential frequencies from the peaks of eigenspectrum as follows:

$$\hat{P}_{EV}=\frac{1}{\sum_{i=P+1}^{M}\frac{1}{\lambda_i}\left|e^H v_i\right|^2} \quad (8)$$

However, with estimated autocorrelations, the EV method differs from MUSIC and produces fewer spurious peaks.

The last method is ESPRIT (Estimation of Signal Parameter via Rotational Invariance Technique) algorithm which allows determining and detecting the parameters of harmonic components with very high accuracy both in frequency and in amplitude estimation independently of the window length. Furthermore, it's a suitable approach to providing reliable spectral results without synchronization effects [20], [21], [22], [23], [24], [33], [38]. It is based on naturally existing shift invariance between the discrete time series which leads to rotational invariance between the corresponding signal subspaces. The eigenvectors U of the autocorrelation matrix of the signal define two subspaces

(signal and noise subspaces) by using two selector matrices Γ_1 and Γ_2.

$$S_1 = \Gamma_1 U,\ S_2 = \Gamma_2 U \tag{9}$$

The rotational invariance between both subspaces leads to the following equation:

$$S_1 = \Phi S_2 \tag{10}$$

Where:

$$\Phi = \begin{bmatrix} e^{j2\pi f_1} & 0 & \cdots & 0 \\ 0 & e^{j2\pi f_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{j2\pi f_M} \end{bmatrix} \tag{11}$$

The matrix Φ contains all information about M components frequencies, and the estimated matrices S can contain errors. Moreover, the TLS (total least-squares) approach finds the matrix Φ as minimization of the Frobenius norm of the error matrix.

Another interesting eigendecomposition method is the minimum norm (MN) algorithm. Instead of forming an eigenspectrum that uses all of the noise eigenvectors. It uses a single vector which is constrained to lie in the noise subspace, and the complex exponential frequencies are estimated from the peaks of the frequency estimation function given by:

$$\hat{P}_{MN}\left(e^{j\omega}\right) = \frac{1}{\left|e^H a\right|^2},\ a = \lambda P_n u_1 \tag{12}$$

The problem, therefore, is to determine which vector in the noise subspace minimizes the effects of the spurious zeros on the peaks of $\hat{P}_{MN}\left(e^{j\omega}\right)$.

It is proposed in this work to apply these methods for detection of different wind turbine generator faults.

VI. Comparative Performance Analysis

The performance (error of estimation) of the subspace methods has been extensively investigated in the literature, especially in the context of the Direction of Arrival (DOA) estimation [33]. In this section, to evaluate the efficiency of the above mentioned fault detectors, with respect to the computation speed, accuracy, degree of frequency estimation dispersion for different level of SNR with a fixed values of fault amplitude.

The faults severity detection is also studied by varying the faults amplitude a_{-1}, a_1 in the interval [0, $0.2a_0$]. The previous frequency estimation methods are applied under different scenarios by simulation in Matlab for a faulty wind turbine generator using 2 pair poles, 4kW/50Hz, 230/400V. The induction generator stator current, showed in figure 1 for a window time of 0.25 s, is simulated by using the signal model described in (1) for the different failure cases described in table I. The parameters of the simulation are illustrated in table II and in table III.

With $f_k(\omega) = f_k(f_0, s, p, k, m)$ is a set of parameters to be estimated at each time n depending on the faults studied cases. Choosing between estimators is a difficult task. Therefore, some quality criteria are needed to determine the best one. The comparison of mean square error (MSE) defined by equation (13) would be helpful for theoretical assessment of accuracy for this purpose.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}_i - f_i\right)^2 \tag{13}$$

$\hat{f}_i$ is the estimated fault frequency

f_i is the exact fault frequency

N is the iterations number

TABLE II. Parameters Used In Simulations

Parameter	Value
s	0,033
L	2
p	2
m	{1,2}
f_0	50 Hz
f_r	29,01 Hz
n_b	12
k	{1,3}
n	1600
F_s	1000 Hz
iterations	200
SNR	[0,100]
Stator Current Amplitude a_0	10 A
Computation Processor	Intel Core2 Duo T6570 2,1 GHz

TABLE III. Faults Simulation Scenarios

	Frequencies (Hz)		Amplitudes (A)		Phase (rad)	
Fault	ϕ_{-1}	ϕ_1	a_{-1}	a_1	φ_{-1}	φ_1
Broken rotor bars	22,53	70,83	1	1	0	0
Inner Bearing damage	89,25	367,74	1	1	0	0
Misalignment	79,01	137,03	1	1	0	0
Air gap eccentricity	74,18	98,35	1	1	0	0

Therefore, for each scenario the fault harmonic amplitude is fixed to $0.1a_0$ as shown in table III. Simulation results for the broken rotor bars fault frequency estimation shows in the figures 2, 3 the evolution of the MSE and the variance average depending on the variation of the SNR. It is very clear that for a stator current having a high level of noise in [0,30]dB, R-MUSIC and ESPRIT gives a nearly identical poor accuracy because these approaches have a high MSE and a high variance values almost constants due to theirs sensibility to noise. Afterwards, Prony and Pisarenko are almost identical and they present a medium accuracy value over other methods, hereafter come Min-Norm then EV method with a constant value of MSE and an increased variation of the variance average.

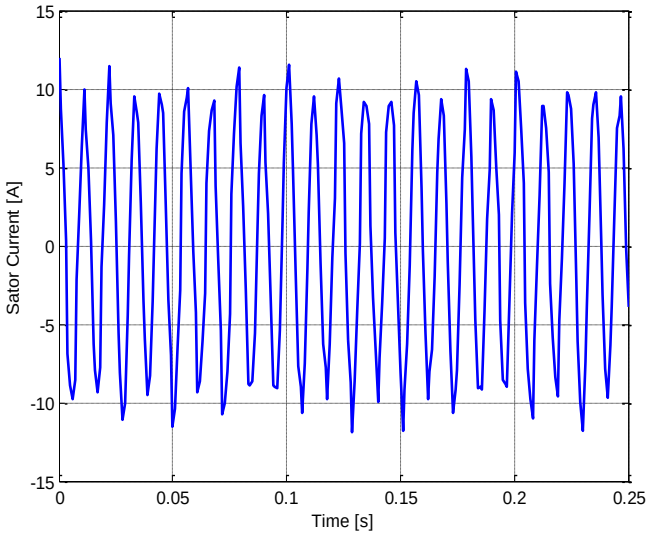


Fig. 1. Faulty induction generator stator current with noise (SNR=30 dB)

Moreover, it is noted that EV gives a good accuracy compared to Min-Norm due to it resistance to noise. Contrariwise, R-MUSIC then ESPRIT becomes more accurate when the SNR increases in [35,100]dB, the accuracy level of these methods exceeds that gives Prony and Pisarenko which have a medium value.

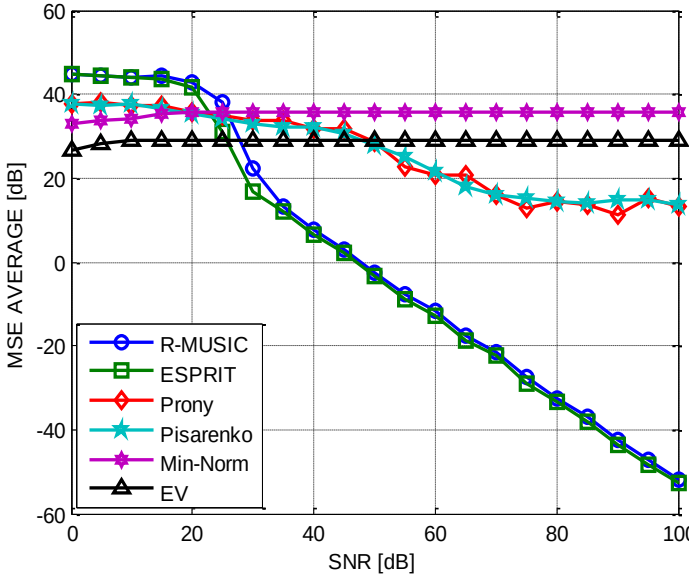


Fig. 2. Mean Square Error Average of Broken rotor bars fault frequency estimation

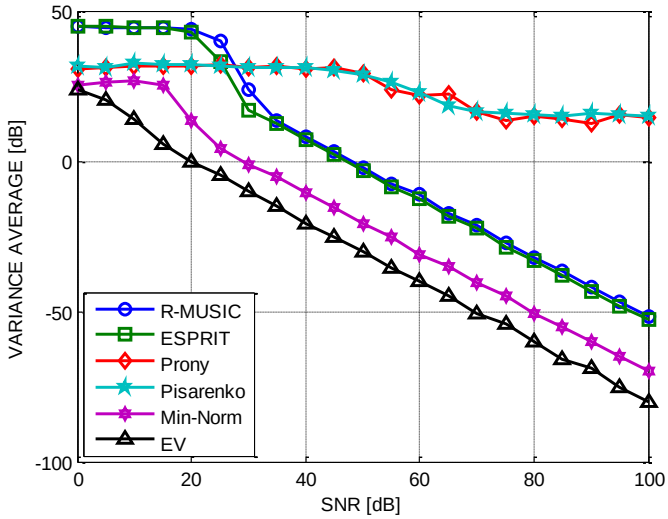


Fig. 3. Variance Average of Broken rotor bars fault frequency estimation

Whereas, figures 4 provide a statistical computation time description, it can be divided into three categories: fast computation methods for Min-Norm, Prony and Pisarenko followed by a medium speed computation for R-MUSIC then EV and ESPRIT come with a high computation time cost. This variation in computation speed for each method can be justified by the autocorrelation and the covariance matrix calculation in addition to the simulated stator current samples size and the sampling frequency used.

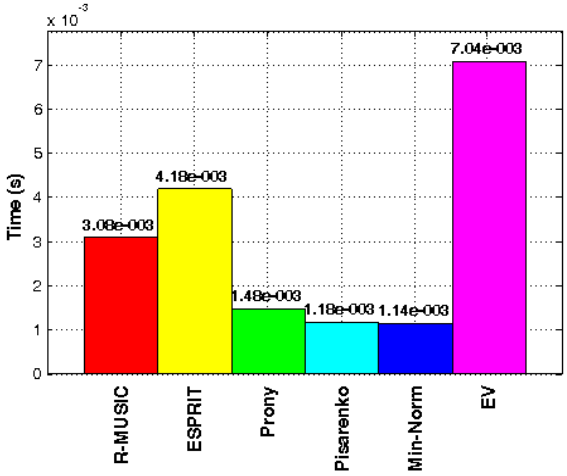


Fig. 4. Computational time cost of average Broken rotor bars fault detection for differents HRM

For inner bearing damage fault frequency estimation, based on the simulation results illustrated in figures 5 and 6, it can be noted that the accuracy of the methods can be classified into three levels: the first level in which ESPRIT and R-MUSIC leading to almost the same very high accuracy even at low SNR because their variance and MSE which decreases rapidly with increasing SNR, in the high level Min-Norm and EV are founded followed by Pisarenko and Prony in the medium accuracy level giving almost constant MSE and variance values.

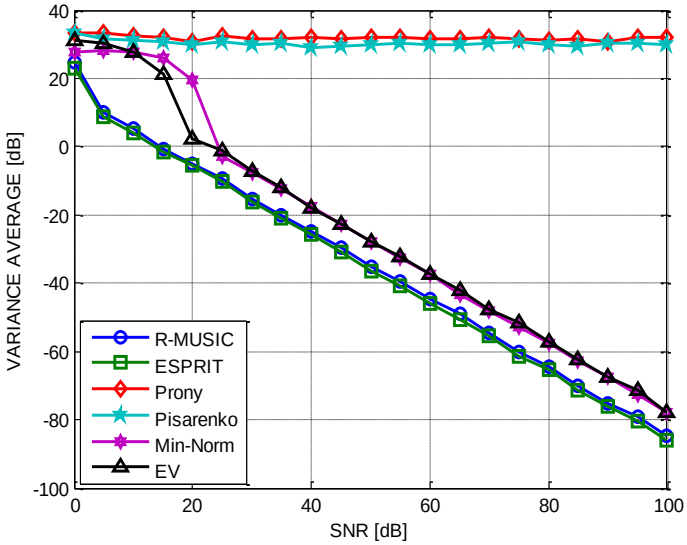


Fig. 5. Variance Average of Inner Bearing damage fault frequency estimation

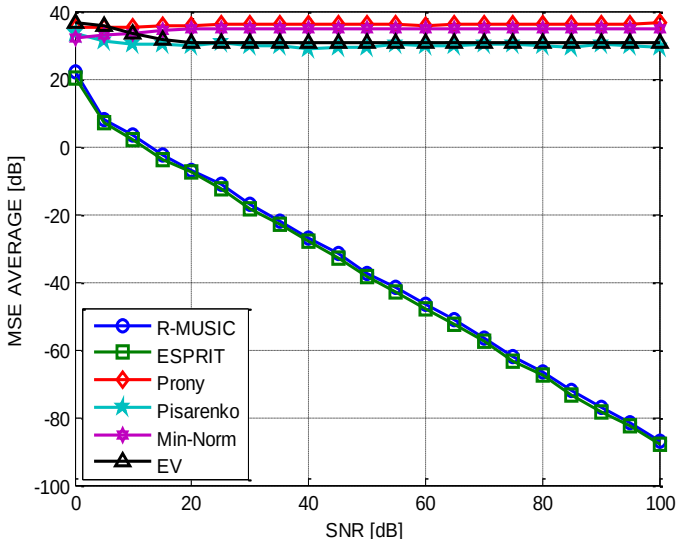


Fig. 6. Mean Square Error Average of Inner Bearing damage fault frequency estimation

This difference in accuracy is mainly due to more disparity in the faults frequency components values. Concerning the computation time, from figure 7, 10 and 13 it is noticed that the studied methods keep the same speed behavior observed in previous scenario.

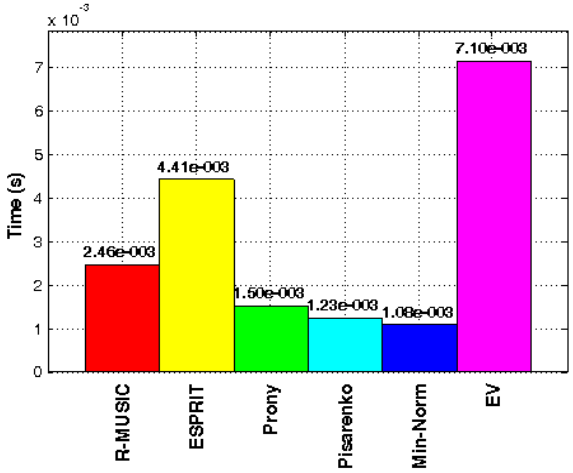


Fig. 7. Computational time cost average of Inner Bearing damage fault detection for differents HRM

Referred to figure 8 and 9, frequency detection precision for misalignment fault is very important and increases significantly by increasing SNR for ESPRIT and R-MUSIC, whereas this accuracy is modest for Prony and Pisarenko mostly beyond an SNR lower than 30dB, after this value EV, Prony and Pisarenko gives almost the same moderate exactitude except Min-Norm which presents a relatively good accuracy which does not reaches ESPRIT and R-MUSIC strictness.

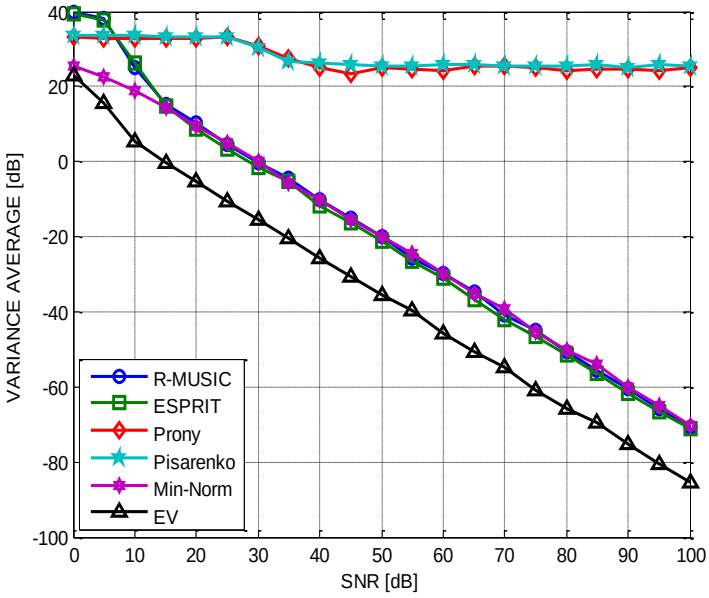


Fig. 8. Variance Average of Misalignment fault frequency estimation

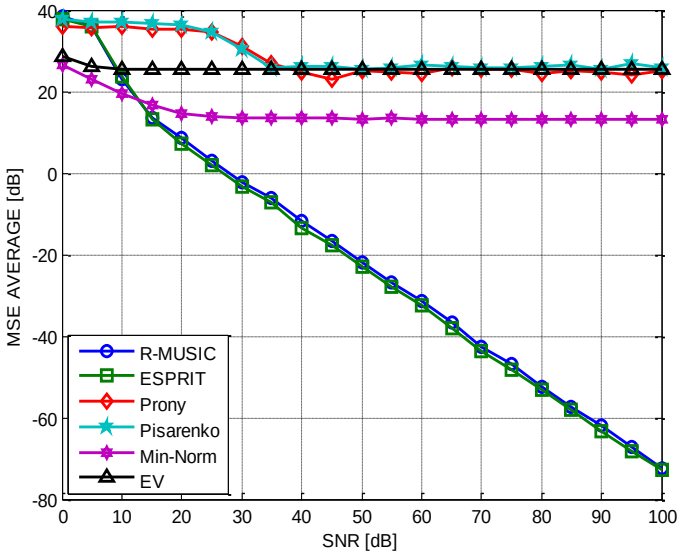


Fig. 9. Mean Square Error Average of Misalignement fault frequency estimation

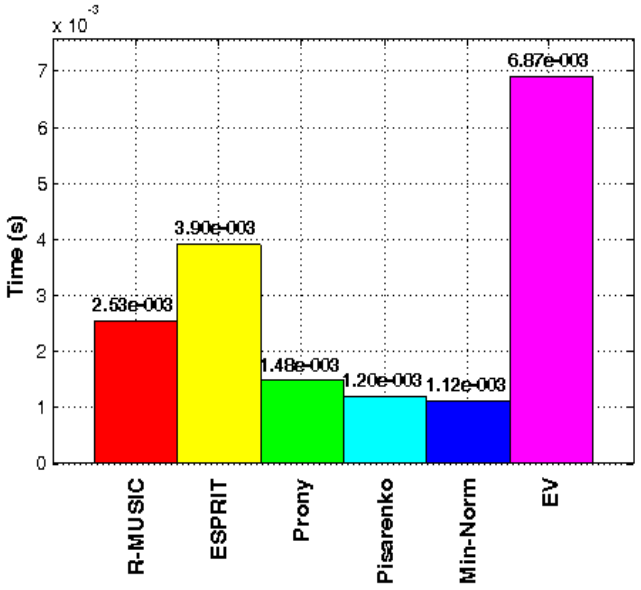


Fig. 10. Computational time cost average of Misalignement fault detection for differents HRM

It seems clearly in figures 11 and 12, for a noisy stator current generator in presence of an air gap eccentricity fault, that Min-Norm and EV are the best choice due to their good accuracy on the contrary of Prony, Pisarenko, R-MUSIC and ESPRIT which gives bad one, but when SNR exceeds 25dB it is observed that R-MUSIC and ESPRIT occupies the first place in accuracy followed by Prony and Pisarenko exceeding Min-Norm for an SNR above 50dB, from this point there are four degrees of quality estimation frequency:

the best one is that take ESPRIT and R-MUSIC, the second one is related to Prony and Pisarenko while others are modest.

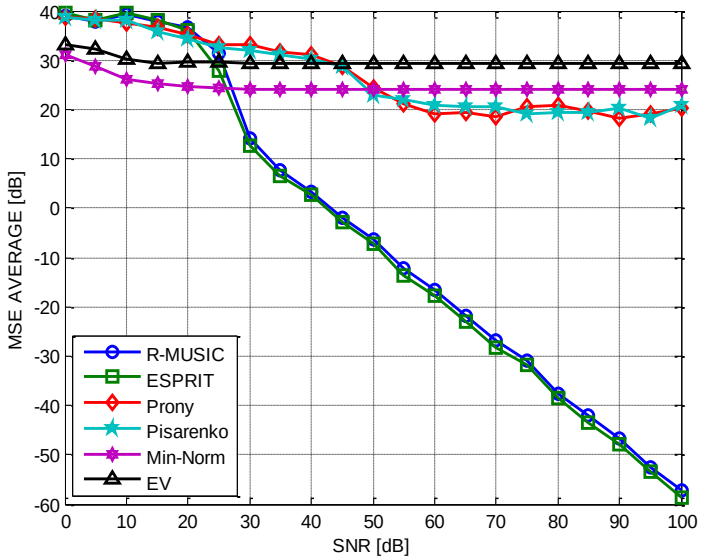


Fig. 11. Mean Square Error Average of Air gap eccentricity fault frequency estimation

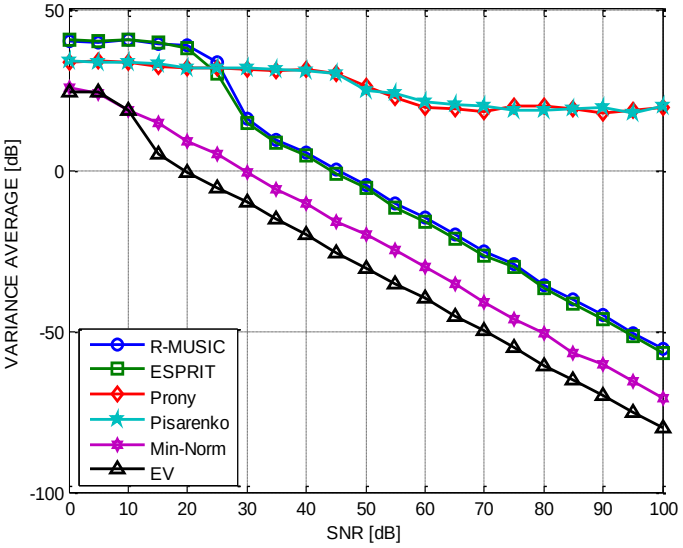


Fig. 12. Variance Average of Air gap eccentricity fault frequency estimation

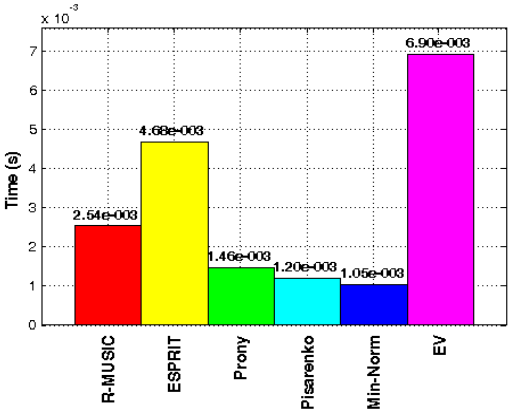


Fig. 13. Computational time cost average of Air gap eccentricity fault detection for differents HRM

To compare and to investigate the ability of the studied methods to detect and to identify clearly the fault frequency components for small amplitudes even in presence of an annoying noise, figures 14, 15, 16 and 17 shows the simulation results obtained for the different studied faults scenarios.

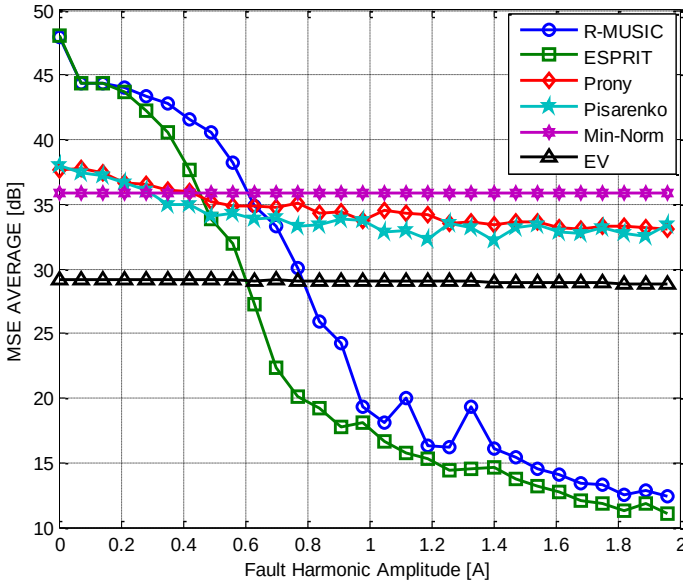


Fig. 14. Mean Square Error Average of Broken rotor bars fault frequency estimation depending on harmonic amplitude variation for differents HRM (SNR=30 dB)

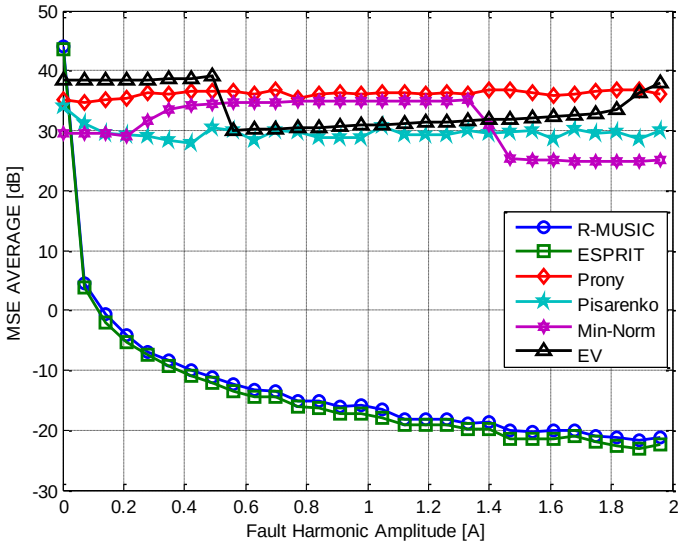


Fig. 15. Mean Square Error Average of Inner Bearing damage fault frequency estimation depending on harmonic amplitude variation for differents HRM (SNR=30 dB)

In a global vision, it is concluded that the ESPRIT and the R-MUSIC methods are very powerful in the detection of fault frequencies despite their corresponding amplitudes are very small, whereas EV and Min-Norm show some instabilities in this identification and their robustness remains limited, on the other hand Prony and Pisarenko have a poor degree of performance compared to ESPRIT and R-MUSIC.

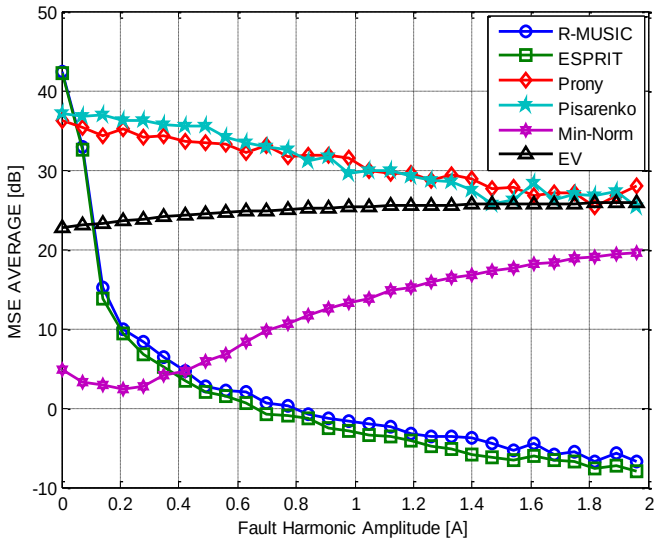


Fig. 16. Mean Square Error Average of Misalignement fault frequency estimation depending on harmonic amplitude variation for differents HRM (SNR=30 dB)

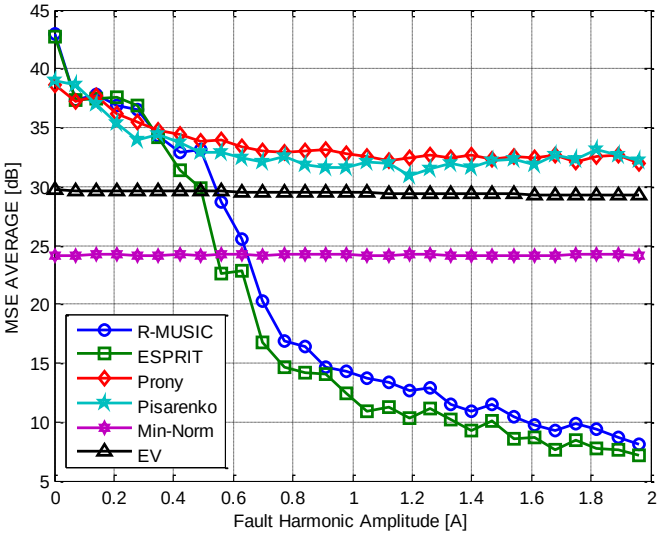


Fig. 17. Mean Square Error Average of Air gap eccentricity fault frequency estimation depending on harmonic amplitude variation for differents HRM (SNR=30 dB)

Generally, ESPRIT and R-MUSIC are even competitive. They have a good detection and resolution capabilities clearly outperform other studied methods. ESPRIT parametric spectral method, accurately and reliably showed high content of fault harmonics in the stator current, justifying their usefulness as a tool for spectral analysis of distorted electric signals in wind power generators although it high computation time cost. Prony and Pisarenko are not very much fruitful when noise increases. They have a limited practical use. EV and Min-Norm are good approaches but sometimes they give a risk of false estimates harmonics due to their roots of eigenvectors which does not correspond to the required frequency. As an outcome, these signal processing methods has been ordered according to the three evaluation criteria previously studied as shown in Table IV

TABLE IV. PERFORMANCE CHARACTERISTICS COMPARAISON OF THE STUDIED PARAMETRIC SPECTRAL METHODS

Method	Time	Accuracy	Risk	Rank
ESPRIT	medium	very high	none	1
R-MUSIC	medium	high	none	2
Min-Norm	small	medium	medium	3
EV	high	medium	medium	4
Pisarenko	small	low	medium	5
Prony	small	low	medium	6

VII. CONCLUSION

In this paper, it has been shown that the high-resolution spectrum estimation methods could be effectively used for wind turbine faults detection which can be achieved by on-line monitoring stator current spectral components produced by the magnetic field anomaly. These techniques aim to separate the observation space in a signal subspace, containing only useful information improving the spectral resolution. An investigation under different conditions is realized to measure robustness and to found efficient tools for detection. The accuracy of the estimation depends on the signal perturbation, fault severity level, the sampling frequency and on the number of samples taken into the estimation process. The comparison has proved the superiority of ESPRIT algorithm than the others followed by R-MUSIC which allows in all cases very high detection accuracy. However, their computation is slightly more complex than the others approaches which can affect their use in real-time implementation. Despite this, ESPRIT can be exploited to design an intelligent embedded system for diagnosis of electromechanical problems in wind turbines generators. As future work, the enhancement of the accuracy and the computation time cost so much more form an important defiance.

REFERENCES

[1] M.C. Mallikarjune Gowda et al, "Improvement of the Performance of Wind Turbine Generator Using Condition Monitoring Techniques", Proceedings of 7th International Conference on Intelligent Systems and Control (ISCO 2013), IEEE 2012

[2] Don-Ha Hwang et al, "Robust Diagnosis Algorithm for Identifying Broken Rotor Bar Faults in Induction Motors", Journal of Electrical Engineering & Technology, JEET, Vol. 9, No. 1, January 2014

[3] Hamid A. Toliyat et al., "Electric Machines Modeling, Condition Monitoring, and Fault Diagnosis", CRC Press Taylor & Francis Group NW 2013, ISBN-13: 978-1-4200-0628-5

[4] M. L. Sin, W. L. Soong and N. Ertugrul, "On-Line Condition Monitoring and Fault Diagnosis – A Survey" Australian Universities Power Engineering Conference, New Zealand, 2003.

[5] Shawn Sheng and Jon Keller et al, "Gearbox Reliability Collaborative Update", NREL U.S. Department of Energy, http://www.nrel.gov/docs/fy14osti/60141.pdf

[6] Shahin Hedayati Kia et al, "A High-Resolution Frequency Estimation Method for Three-Phase Induction Machine Fault Detection", IEEE Transactions on Industrial Electronics, Vol. 54, No. 4, AUGUST 2007.

[7] Ouadie Bennouna et al, "Condition Monitoring & Fault Diagnosis System for Offshore Wind Turbines", https://zet10.ipee.pwr.wroc.pl/record/425/files/invited_paper_3.pdf

[8] Elie Al-Ahmar et al, "Wind Energy Conversion Systems Fault Diagnosis Using Wavelet Analysis", International Review of Electrical Engineering 3, 4 (2008) 646-652, http://hal.univ-brest.fr/docs/00/52/65/07/PDF/IREE_2008_AL-AHMAR.pdf

[9] M.L. Sin et al, "Induction Machine On-Line Condition Monitoring and Fault Diagnosis – A Survey", http://www.academia.edu/416441/Induction_Machine_on_Line_Condition_Monitoring_and_Fault_Diagnosis_A_Survey

[10] K. K. Pandey et al, "Review on Fault Diagnosis in Three-Phase Induction Motor", MEDHA – 2012, Proceedings published by International Journal of Computer Applications® (IJCA)

[11] Niaoqing Hu et al, "Early Fault Detection using A Novel Spectrum Enhancement Method for Motor Current Signature Analysis", 7th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED'08), University of Cambridge, UK, Feb 20-22, 2008

[12] E. Al Ahmar et al, "Advanced Signal Processing Techniques for Fault Detection and Diagnosis in a Wind Turbine Induction Generator Drive Train: A Comparative Study", IEEE Energy Conversion Congress and Exposition (ECCE), 2010, Atlanta : États-Unis (2010)

[13] El Houssin El Bouchikhi et al, "A Comparative Study of Time-Frequency Representations for Fault Detection in Wind Turbine", IECON 2011 - 37th Annual Conference on IEEE Industrial Electronics Society

[14] Francisco José Vedreño Santos, "Diagnosis of Electric Induction Machines in Non-Stationary Regimes Working in Randomnly Changing Conditions", Thesis Report, Universitat Politècnica de València, November 2013

[15] M. Hayes, "Digital signal processing and modeling", Wiley, New York, NY, 1996.

[16] T.T. Georgiou, "Spectral Estimation via Selective Harmonic Amplification", IEEE Trans. on Automatic Control, 46(1): 29-42, January 2001.

[17] T.T. Georgiou, "Spectral analysis based on the state covariance: the maximum entropy spectrum and linear fractional parameterization," IEEE Trans. on Automatic Control, 47(11): 1811-1823, November 2002.

[18] P. Billingsley, "Probability and Measure", second edition, John Wiley and Sons, New York, 1986.

[19] V.F. Pisarenko, "The Retrieval of Harmonics from a Covariance Function", Geophysics J. Roy. Astron. Soc. 33 (1973), 347-366.

[20] R. Roy and T. Kailath, "ESPRIT - Estimation of Signal Parameters via Rotational Invariance Techniques", IEEE Transactions on Acoustics, Speech, and Signal Processing. ASSP-37 (1989), 984-995.

[21] R.O Schmidt, "A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation", Ph.D. thesis, Stanford University, Stanford, CA, 1981.

[22] Sophocles J. Orfanidis, "Optimum Signal Processing", MCGRAW-HILL publishing company, New York, ny, 2nd edition 2007.

[23] John L. Semmlow, "Biosignal and Biomedical Matlab-Based Applications", Marcel Dekker, Inc New York 2004.

[24] Gérard Blanchet and Maurice Charbit, "Digital Signal and Image Processing using Matlab", ISTE USA 2006.

[25] Yassine Amirat et al, "Wind Turbine Bearing Failure Detection Using Generator Stator Current Homopolar Component Ensemble Empirical Mode Decomposition", IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society

[26] El Houssin El Bouchikhi et al, "A Parametric Spectral Estimator for Faults Detection in Induction Machines", Industrial Electronics Society, IECON 2013 - 39th Annual Conference of the IEEE

[27] Mohamed Becherif et al, "On Impedance Spectroscopy Contribution to Failure Diagnosis in Wind Turbine Generators", International Journal on Energy Conversion 1, 3 (2013) pages 147-153.

[28] Ioannis Tsoumas et al, "A comparative study of induction motor current signature analysis techniques for mechanical faults detection, SDEMPED 2005 - International Symposium on Diagnostics for Electric Machines", Power Electronics and Drives Vienna, Austria, 7-9 September 2005

[29] Young-Woo Youn et al, "MUSIC-based Diagnosis Algorithm for Identifying Broken Rotor Bar Faults in Induction Motors Using Flux Signal, Journal of Electrical Engineering & Technology", JEET, Vol. 8, No. 2, 2013, pages 288-294.

[30] El Houssin El Bouchikhi et al, "Induction Machine Fault Detection Enhancement Using a Stator Current High Resolution Spectrum", IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society

[31] Yassine Amirat et al, "Wind Turbines Condition Monitoring and Fault Diagnosis Using Generator Current Amplitude Demodulation", IEEE International Energy Conference and Exhibition (EnergyCon), 2010

[32] Neelam Mehala et al, "Condition monitoring methods, failure identification and analysis for Induction machines", International Journal of Circuits, Systems and Signal Processing, Issue 1, Volume 3, 2009, pages 10-17

[33] Zbigniew Leonowicz, "Parametric methods for time-frequency analysis of electric signals", Politechnika Wrocławska Wroclaw University of Technology, Poland, 2nd edition January 2007.

[34] Przemyslaw Janik et al, "Advanced Signal Processing Methods for Evaluation of Harmonic Distortion Caused by DFIG Wind Generator", 16th PSCC, Glasgow, Scotland, July 14-18, 2008.

[35] H. C. So et al, "Linear Prediction Approach for Efficient Frequency Estimation of Multiple Real Sinusoids: Algorithms and Analyses", IEEE Transactions on Signal Processing, Vol. 53, No. 7, July 2005

[36] Tavner, P. "How Are We Going to Make Offshore Wind Farms More Reliable?", SUPERGEN Wind General Assembly on March 20, 2011 at Durham University, UK.

[37] Sheng, S. "Investigation of Various Wind Turbine Drive train Condition Monitoring Techniques", Wind Turbine Reliability Workshop, August 2–3, 2012 Albuquerque, NM.

[38] André Quinquis, "Digital Signal Processing using MATLAB", ISTE Ltd, London UK, 2008

[39] Łobos T. et al., "Advanced signal processing methods of harmonics and interharmonics estimation", IEE Seventh International Conference on Developments in Power System Protection, Amsterdam, 9–12 April 2001, pp. 315–318.

A Survey of Unstructured Text Summarization Techniques

Sherif Elfayoumy
School of Computing
University of North Florida
Jacksonville, Florida

Jenny Thoppil
School of Computing
University of North Florida
Jacksonville, Florida

***Abstract*—Due to the explosive amounts of text data being created and organizations increased desire to leverage their data corpora, especially with the availability of Big Data platforms, there is not usually enough time to read and understand each document and make decisions based on document contents. Hence, there is a great demand for summarizing text documents to provide a representative substitute for the original documents. By improving summarizing techniques, precision of document retrieval through search queries against summarized documents is expected to improve in comparison to querying against the full spectrum of original documents.**

Several generic text summarization algorithms have been developed, each with its own advantages and disadvantages. For example, some algorithms are particularly good for summarizing short documents but not for long ones. Others perform well in identifying and summarizing single-topic documents but their precision degrades sharply with multi-topic documents. In this article we present a survey of the literature in text summarization. We also surveyed some of the most common evaluation methods for the quality of automated text summarization techniques. Last, we identified some of the challenging problems that are still open, in particular the need for a universal approach that yields good results for mixed types of documents.

Keywords*—*text summarization; unstructured data; text mining; unstructured data analytics

I. Introduction

The rapid growth of online information services, social media and other digital format documents means that huge amounts of information are becoming immediately available and readily accessible to a large number of end-users. However, human ability to organize and understand a large number of documents is limited. This well-known information overload problem is most acute when we need to make a decision or understand something deeply, which typically involves reviewing several documents, but have limited time. Reading through long documents consumes precious time in understanding the gist of the document.

Web search engines look for documents from the Internet based upon user supplied queries. They not only overwhelm users with too many results, they also provide documents that may not be very relevant to the topic being studied by the user. For example, if the user is searching using some keyword and the search engine finds it somewhere inside a document, that document will be a "search hit" even if the document is not really relevant to the keyword. The most common search method is based on maintaining an inverted list (text index) of documents' text. Not only precision is hurt by indexing every word in the document, excluding stop words, but also efficiency is adversely impacted. If summaries are indexed and searched instead, index size will be considerably smaller and search hits will be of better quality (fewer false positives) [1]. This can be explained using the definition of Precision and Recall measures used in information retrieval. Precision is defined as the percentage of the relevant items in the returned set and Recall is the percentage of the relevant items in the returned set compared to those in the collection. If the whole collection is retrieved, then the Recall is 100%, but Precision is low. Most search engines suffer from this problem (high Recall and low Precision). If search engines search only a document's primary ideas, instead of every word, then Recall will likely not be decreased but Precision will likely improve. Hence, an automated facility for summarizing documents to improve productivity is desirable. A good summarization system should include only sentences that are most important to a document's theme; it must also cover all documents' topics [2].

Using a summary instead of the whole document as a representative of what the document is about would mean processing a fraction (20% or less) of the document's text, yet yield better Precision and lesser processing time for search queries. In order to determine the requirements of a good summarization system, many text summarization approaches were reviewed. An in-depth review of text summarization literature was conducted and results from this study along with a description of each algorithm, its strengths and weaknesses are presented in this article. Section II presents an overview of the major types of text summarization techniques. Section III provides detailed information on unsupervised text summarization techniques. The evaluation techniques used for assessing the quality of text summarization systems are also discussed in section IV. It was found that due to the shortcomings of the text summarization approaches currently available, there is a lack of a universal approach for document summarization that provides high Precision and Recall with various types of text corpora.

II. Text Summarization by Classification

Many research papers and books related to natural language processing and computational linguistics were thoroughly investigated in order to determine current techniques used for automated text summarization and in particular their

advantages and disadvantages. Text summarization techniques were classified by Hahn and Mani [3] as follows:

A. Query-relevant Summarization

A query-relevant summary presents the document's contents that are closely related to an initial search query. This can be achieved by extending conventional information retrieval technologies. Depending on the user's supplied query, the text documents are searched for matches with that query, and a summary is created on the fly, which contains the sentences that have the query matches.

The selection of sentences based on their ranking, with respect to a query, using latent semantic analysis (LSA) was proposed by Gong and Liu [2]. Park et al. proposed a new approach using a combination of Non-negative Matrix Factorization and K-means clustering to identify sentences based on a query. Their approach produced better performance than LSA [4]. Tang et al. retrieve relevant documents to a query, use a unified probabilistic approach to discover query-oriented topics and apply four scoring methods to calculate the importance of each sentence. Sentences with the highest score make the summary of each document [5].

B. Generic Summarization

A generic summary provides an overall sense of the document's contents. It contains the main topics of the document, while keeping redundancy to a minimum. As neither query nor topic is provided to the summarization process, it is challenging to develop a high quality generic summarization method [2]. Generally, text summary extraction from a document can be done using one or more of the following approaches:

a) Sentence Extraction

In this method, original pieces from the source document are selected and concatenated to yield a shorter text. This technique is easy to adapt to large sources of data. A Conditional Random Field (CRF) framework was proposed by Shen et al. In their framework, the summarization problem is viewed as a sequence labeling problem where a document is a sequence of sentences that are labeled as 1 or 0 based on the label assignment to other sentences [6]. Daume and Marcu presented BAYESUM which is a Bayesian Summarization model for query expansion. This model was found to be work well in purely extractive settings [7].

b) Sentence Abstraction

This method paraphrases in more general terms what the text is about. This is done using very sophisticated algorithms. It is easy to adapt to higher compression rates [3]. Kinght and Marcu presented corpus-based methods for attacking the sentence abstraction problem, one using the noisy-channel framework, and other using a decision-based model. While most corpus-based work focuses on keyword extraction, this work focused on constructing new whole sentences by analyzing existing, manually produced, compressions [8].

c) Supervised Approaches

These approaches make use of human-made summaries or extracts to identify features or parameters of summarization algorithms. In these methods, a human user decides which parameters are important for text summary and accordingly the summary is generated. Bravo-Marquez and Manriquez trained ranking functions using linear regressions and ranking SVMs, which are also combined using Borda count [9]. Top ranked sentences are concatenated and used to build summaries, which are compared with the first sentences of the distant summary using ROUGE evaluation measures [10]. Experimental results obtained showed that the combination of different ranking techniques improves the quality of the generated summary.

d) Unsupervised Approaches

These approaches determine the relevant parameters without regard to human–made summaries [11]. The summary is generated without any user input. Probabilistic Latent Semantic Indexing (PLSI) is an unsupervised learning method based on statistical latent class models. PLSI was applied to document clustering by Hoffman [12]. In contrast to standard

Latent Semantic Indexing (LSI) by Singular Value Decomposition, the probabilistic variant, PLSI, has a solid statistical foundation and defines a proper generative data model. Retrieval experiments indicated substantial performance gains over LSI. PLSI was further developed into a more comprehensive Latent Dirichlet Allocation (LDA) model by Blei et al [13]. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. Topic probabilities provide an explicit representation of a document.

The unsupervised approaches do not require user input in deciding the important features of the document, requiring a more sophisticated algorithm to compensate for lack of human intervention. We believe unsupervised summaries provide a higher level of automation which makes them more suitable for processing Big Data.

III. UNSUPERVISED GENERIC TEXT SUMMARIZATION

In this section we investigate further generic text summarization using unsupervised approaches for sentence extraction. Generic text summary can improve the processing time and precision of information retrieval, since the summary has to be created only once and contains the most important themes of the document. In contrast, query-related summaries need to be created every time a query is provided by the user. Moreover, it is possible the summaries do not have the query as the main topic of the documents retrieved. The sentence extraction approach is a simpler but effective way of extracting main themes of the document as compared to sentence abstraction, which involves many complicated linguistic and natural language processing algorithms that require a lot of processing time. The following generic unsupervised text summarization algorithms have been amongst the most prominent in the literature.

A. Cosine Similarity

The vector space model using cosine measure is one of the most widely used models for text retrieval, mainly because of its conceptual simplicity. Sentences and queries are represented in a high-dimensional space, in which each dimension of the space corresponds to a word in a sentence collection [14]. The

most relevant sentences for a query are expected to be those represented by the vectors closest to the query.

This method can be slightly modified to calculate a weight for each sentence with respect to its relevance to the entire document. In order to calculate the cosine measure of a sentence, the frequency of each term in the entire document (*docfreq*) and the frequency of the term in a particular sentence *(termfreq*) are calculated. Then for each sentence, i.e. query, the cosine angle between the query and the entire document is calculated using the formula below. If the cosine measure is highest, i.e. the cosine angle between query and document is smallest, then that sentence is the most relevant to the document. Thus the sentences are ranked according to their cosine measures and a summary is created using top ranked sentences. The formula for cosine similarity is as follows:

$$\cos(\text{termfreq},\text{docfreq}) = \frac{\sum_{i=1}^{n} termfreq.docfreq}{\left(\sqrt{\sum_{i=1}^{n} termfreq^2} \times \sqrt{\sum_{i=1}^{n} docfreq^2}\right)}$$

where *n* = number of terms per sentence

The Cosine Similarity technique is not well-suited for obtaining diverse topics in a document, although it does an excellent job of selecting the most relevant sentences in the document. In Maximal Marginal Relevance (MMR), the Cosine Similarity technique is changed to add diversity to the document summary [15].

B. Relevance Measure

Gong and Liu [2] proposed a relevance measure algorithm, which is also based on ranking sentences using their relevance scores. This algorithm works as follows: The weighted frequency vector is obtained for each sentence using the local weight of each term and its global weight over the document, where each term's weight is obtained as

$$aji = L(tji).G(tji)$$

where *L(tji)* is the local weight for term j in passage i and *G(tji)* is the global weight for term j.

Vector length normalization, also referred to as cosine normalization, is carried out and the weight of each sentence is obtained. The sentence with highest relevance score is extracted and added it to summary. All the terms contained in the sentence are deleted from the original document. The sentence itself is deleted and weighted term frequency vector for the document is recomputed. Again sentence with highest relevant score is found and this process is continued until the number of sentences in the summary reaches a predefined value [2].

C. Latent Semantic Analysis using SVD

Singular value decomposition (SVD) is a method of word co-occurrence analysis using a dimensionality reduction approach. In the process of dimensional reduction, co-occurring terms are mapped onto the same dimensions in the reduced space, thus increasing similarity in the representation of semantically similar sentences [15]. In this method, the weight of the sentences is first obtained using the same principle as described in [2] and then a sentence matrix A= [A1, A2, …An] with each column vector Ai representing the weighted term vector of sentence i is created. If there are m terms and n sentences, then matrix A is of dimension m*n. Using singular value decomposition, A= USV^T, where the columns of U (m*dimension) are left singular vectors, S (dimension * dimension) gives the non-negative singular values, and V^T (dimension *n) columns are right singular vectors. The first right singular vector is selected and the sentence with the largest index value is selected and included in the summary [2]. The next right singular vector representing the next dimension is selected and the largest index valued sentence is added to the summary. Thus, this method chooses sentences from every dimension covering all topics in the document.

In Enhanced Latent Semantic Analysis using SVD [16], for each sentence vector in matrix V, its components are multiplied by corresponding singular values, to compute each sentence length. The reason for using the multiplication is to favor the index values in the matrix V that correspond to the highest singular values; i.e. the most significant topics. The sentence weight is calculated as follows:

$$S_k = \sqrt{\sum_{i=1}^{n} V_{k,i}^2 \times S_i^2}$$

where S_k is the sentence with sentence number k and n = number of dimensions.

The Latent Semantic Analysis using SVD, though a good dimensionality reduction technique, has two disadvantages. It is necessary to use the same number of dimensions as the number of sentences chosen for a summary. If a high number of dimensions of the reduced space is chosen, the probability of selecting a significant topic in the summary is reduced. Hence, it may not give the most relevant sentences for longer documents. Also, sentences with large sentence weights, but not the largest (they do not win in any dimension), will not be chosen although contents may be very suitable for the summary [16]. Hence in Enhanced Latent Semantic Analysis technique, the weight of each sentence is further calculated with respect to the entire document, not just with respect to each dimension, so that sentences can be correctly ranked.

D. Maximal Marginal Relevance (MMR)

MMR is based on the vector space model of text retrieval [15][17] and is well suited for query-based and multi-document summarization. It chooses sentences according to a weighted combination of their relevance to a query and their redundancy with sentences that have already been extracted using Cosine Similarity. The MMR score $S_{CMMR(i)}$ for a given sentence S_i in a document is given by

$$S_{CMMR(i)} = [\lambda Sim(S_i, D) - (1-\lambda)\max(S_i, Summ)]$$

where *D* is the average document vector, *Summ* is the average vector from the set of sentences already selected, and λ trades off between relevance and redundancy. *Sim* is the cosine similarity between the two documents.

When λ=1, it computes the incrementally standard relevance ranked list. When λ=0, it computes a maximal

diversity ranking among the documents. When MMR was compared with Enhanced LSA, MMR yielded better Precision [17]. The Maximal Marginal Relevance measure is commonly used for multi-document summarization.

E. Full Coverage Summarizer

The first phase in the Full Coverage Algorithm is to parse a document into sentences [18]. During this phase, stop-words are removed and the Porter stemming algorithm is applied to stem the words in the document to their base forms. The entire document is then treated as a query to each individual sentence. The second step is to calculate the subset of sentences that cover the entire concept space of the document. The highest ranked sentence is selected using Cosine Similarity. The words that appear in the highest ranked sentence are removed from the query and the process is repeated until no words can be removed from the query, thus obtaining the summarized document. Mallett et al. also compared the Full-Coverage summarizer with MEAD and found that the Full-Coverage summarizer outperforms the MEAD clustering technique [18].

F. MEAD

MEAD is a multi-document summarizer which generates summaries using cluster centroids produced by topic detection and tracking system (TDT) [19]. MEAD uses the online document clustering system, CIDR, to produce the clusters and then uses its own weighting scheme to rank the sentences in the cluster. The CIDR algorithm initially places the first document by itself in the first cluster. The centroids of the cluster are a group of words that represent a cluster of documents. When new sentences are processed, they are compared with the centroids of the existing cluster. Centroids of a cluster are the weighted averages of the *tf*idf* values of the documents already assigned in the cluster, where *tf = frequency of term* and *idf = inverse document frequency*.

Similarity between a document and a centroid is measured using the cosine (normalized inner product) of the corresponding *tf*idf* vector. If the similarity goes below a predefined threshold value, a new cluster is created.

Centroid-based summarization (CBS) uses the centroids of the clusters produced by CIDR to identify sentences central to the topic of the entire cluster. MEAD combines the following three parameters to find the score of a sentence within each cluster:

1) Centroid value – The centroid value of sentence Si is computed as the sum of the centroid values Cw,i of all the words in the sentence.

$$Ci = \sum C_{w,i}$$

2) Positional Value – The first sentence in a document gets the same score Cmax as the highest-ranking sentence in the document using the centroid value. The score for all the sentences within the document is computed as:

$$P_i = (n - i + 1)/n \times C_{\max}$$

3) First sentence overlap – Overlap value is computed as the inner product of the sentence vectors for the current sentence i and the first sentence of the document.

$$F_i = S_1 \times S_i$$

4) Redundancy Penalty -

$$Rs = 2 \times \left(\frac{noofoverlappingwords}{noofwords_{sentence1} \times noofwords_{sentence2}} \right)$$

$$Score(S_i) = w_c C_i + w_p P_i + w_f F_i + w_r R_s$$

Using this score, the sentences are ranked and chosen from each cluster in MEAD.

G. K-means Clustering Followed by tf.idf

A modified K-means algorithm using the Minimum Description Length Principle (MDL) is used, where the number of clusters are estimated, which otherwise has to be supplied by the user [20]. Using K-means, the diversity in the document is obtained in the form of clusters. After clusters are identified, sentences in each cluster are ranked based on the *tf*idf* value, where *tf = term frequency* of each term and *idf = inverse document frequency*, using term frequency over the entire document (*doc*) and the weight of each sentence:

$$W_s = \sum_{x=1}^{n} \left(1 + \log(tf(x).idf(x))\right)$$

where n = number of terms per sentence,

$idf(x) = log\ (N/\ doc(x))$ where N = number of sentences

The weighting scheme is obtained to reduce the redundancy in the document and to choose the sentence with largest weight in the summary. Thus, one or more sentences are chosen from each cluster and added into the summary.

After reviewing the above algorithms, it was clear that each works well given some assumptions, but they do not fulfill all requirements in all circumstances. For example, Cosine Similarity is a good and simple algorithm, if the same words are used for explaining a certain situation. In such cases, it will give very good results. But if the same words are not repeated in the document for a particular context, its Precision is much reduced.

Enhanced Latent semantic analysis using SVD does a good job in finding co-occurrence of terms in a document. It is, therefore, able to find diverse topic areas in the document, but as the number of sentences in the document increases, its Precision drastically degrades, since the number of dimensions in the vector space increases. MMR is a good multi-topic summarizer, but it is not very effective for single-topic documents. Clustering techniques, MEAD, and K-means Clustering are time consuming.

IV. Text Summarization Evaluation Techniques

Objectively evaluating the quality of summarizers is not an easy task, because there are various evaluation metrics. Moreover, arguably there is no "ideal" summary to compare against [21]. Typically, the base-line is a summary generated by a human being. The commonly used metrics include Precision, Recall, Kappa, Relative Utility and n-grams. They are used to compare the automated summary against the manually produced summary.

A. Precision and Recall

Using Precision and Recall measures may be the simplest, but most effective evaluation technique used in text summarization. Precision is defined as the percentage of relevant sentences in the returned set and Recall is the percentage of the relevant sentences in the collection that are in the returned set [14]. $Sum_{manual} \cap Sum_{automated}$ is the set of sentences selected by both automated summarizer and manual summarizer where Sum_{manual} is the set of sentences selected by manual summarizer and $Sum_{automated}$ are the sentences selected by the automated summarizer. Then Precision and Recall are calculated as follows:

$$\Pr ecision = \frac{Sum_{manual} \cap Sum_{automated}}{Sum_{automated}}$$

$$\mathrm{Re}call = \frac{Sum_{manual} \cap Sum_{automated}}{Sum_{manual}}$$

Normally there is more than one judge for summarizing a document manually, and the common sentences among the judges need to be taken as relevant sentences. The amount of agreement between the manual and automated summaries is an important factor in calculating Precision and Recall metrics. A drawback of using Precision and Recall only for evaluating summarizers is that agreement may be by chance and the Precision and Recall approach does not take chance agreement into account [21].

B. Kappa Coefficient

Kappa is an evaluation measure which is increasingly used in NLP (Natural Language Processing) research. It factors out random agreement that Precision and Recall measures do not. Random agreement is defined as the level of agreement which would be reached by random annotation using the same distribution of categories as real annotators [21]. The Kappa coefficient (K) measures pair wise agreement among a set of judges making category judgments and is computed as follows:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where *P(A)* is the probability that the judges agree and *P(E)* is the probability of which judges are expected to agree by chance [22].

Using the Kappa Coefficient along with Precision and Recall gives an accurate evaluation of how well an automated summarizer performs compared to a manual summarizer.

C. Relative Utility

Relative Utility (RU) is a measure for evaluating extractive summarizers. RU is applicable in both single-document and multi-document summarization. When the target sentences are given, the judges (manual and automated summarizers) pick different sentences. This is called Summary Sentence Substitutability (SSS) [23].

RU agreement is defined as the relative score that one judge would get, given his own extract and the other judge's sentence judgments. In RU, a number of judges are asked to assign utility scores to all n sentences in a document.

The top e sentences according to utility score are then used as a sentence extract of size e.

In situations where automated summaries are compared to manual summaries where sentences are not ranked, the Relative Utility technique could not be used as an evaluation technique.

D. BLEU and n-grams

The main idea of the BLEU (Bilingual Evaluation Understudy) method is to measure the translation closeness between a candidate machine translation and a set of reference human translations with a numerical metric. In the unigram precision model, the precision is calculated by simply counting the number of candidate translation words (unigrams) which occur in any reference translation and then divide by the total number of words in the candidate translation [24]. Machine translation system can over-generate reasonable words; hence, the modified unigram technique first counts the maximum number of times a word occurs in single reference translation. Then the total count of each candidate word is clipped by its maximum reference count, the clipped counts are added and then divided by the total (unclipped) number of candidate words. The modified n-gram precision is computed similarly for any n.

The formula for modified n-gram precision on a block of text is as follows:

$$Pn = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in \{Candidates\}} \sum_{n-gram` \in C`} Count(n-gram`)}$$

The BLEU technique is applicable only in situations where automated machine translations are performed.

V. Conclusion

The document summarization problem is an interesting problem due to its impact on information retrieval methods as well as on the efficiency of decision making processes, particularly in the age of Big Data. Although a wide variety of text summarization techniques and algorithms have been developed there is a need for new approaches to produce precise and reliable document summaries that can tolerate differences in document characteristics.

We plan to use the best of bread among the existing techniques to create an ensemble that is capable of producing superior results on mixed document corpora.

References

[1] Walters, William H. "Comparative Recall and Precision of Simple and Expert Searches in Google Scholar and Eight Other Databases," portal: Libraries and the Academy, vol. 11, no. 4, 2011, pp. 971-1006.

[2] Gong, Y. and X. Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," Proceedings of the 24th annual international ACM/SIGIR conference on research and development, ACM-Press, New Orleans, LA, 2001, pp. 75-95.

[3] Hahn, U. and I. Mani, "The Challenges of Automatic Summarization," IEEE Computer, vol. 33, no. 11, 2000, pp. 29-36.

[4] S. Park, J. Lee, D. Kim, and C. Ahn, "Multi-document Summarization Based on Cluster Using Non-negative Matrix Factorization," Lecture Notes in Computer Science, vol. 4362, pp. 761-770, 2007.

[5] J. Tang, L. Yao, and D. Chen, "Multi-topic Based Query-oriented Summarization," Proceedings of SIAM International Conference on Data Mining (SDM), Sparks, NV, 2009.

[6] D. Shen, J. Sun, H. Li, Q. Yang, and Z. Chen, "Document Summarization Using Conditional Random Fields," Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), 2007, pp. 2862–2867.

[7] H. Daumé and D. Marcu, "Bayesian Query-focused Summarization," Proceedings of the 21st International Conference on Computational Linguistics, 2006, pp. 305-3012.

[8] K. Knight, and D. Marcu, "Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression," Artificial Intelligence Journal, vol. 139, no. 1, 2002, pp. 91–107.

[9] Bravo-Marquez, Felipe and Manriquez, Manuel, "A Zipf-Like Distant Supervision Approach for Multi-document Summarization Using Wikinews Articles," Lecture Notes in Computer Science, vol. 7608, 2012, pp. 143-154.

[10] Lin and Chin-Yew, "ROUGE: A Package for Automatic Evaluation of Summaries," Proceedings of the 19th International Conference on Computational Linguistics, Barcelona, Spain, 2004, pp. 74-81.

[11] H. Zha, "Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering," Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2002, pp. 113-120.

[12] T. Hoffman, "Probabilistic Latent Semantic Indexing," Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 1999, pp. 50-57.

[13] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[14] C. Manning, and H. Schlutze, "Foundations of Statistical Natural Language Processing," The MIT Press, 1999.

[15] J. Carbonell, and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," Proceedings of the 21st Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998, pp. 335-336.

[16] J. Steinberger and K. Jezek, "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation," Proceedings of the International Conference on Information Systems Implementation and Modeling (ISIM) 2004, pp. 93-100.

[17] G. Murray, S. Renals and J. Carletta, "Extractive Summarization of Meeting Recordings," Proceedings of Interspeech, Lisbon, Portugal, 2005, pp. 593-596.

[18] D. Mallett, J. Elding, and M. Nascimento, "Information-content Based Sentence Extraction for Text Summarization," Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC), vol. 2, pp. 214, 2004.

[19] D. Radev, H. Jing, and M. Budzikowska, "Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies," Proceedings of the NAACL-ANLP Workshop on Automatic Summarization, vol. 4, pp. 21-30, 2000.

[20] T. Nomoto, and Y. Matsumoto, "A New Approach to Unsupervised Text Summarization," Proceedings of the 24th International Conference on Research in Information Retrieval (SIGIR), pp. 26–34, 2001.

[21] D. Radev, E. Hovy, and K. McKeown, "Introduction to the Special Issue on Summarization," Computational Linguistics, 28(4), pp. 399-408, 2002.

[22] J. Carletta, "Assessing Agreement on Classification Tasks: The Kappa Statistic," Computational Linguistics, vol. 22, no. 2, pp. 249-254, 1996.

[23] D. Radev and D. Tam, "Summarization Evaluation Using Relative Utility," Proceedings of the 12th International Conference on Information and Knowledge Management, pp. 508-511, 2003.

[24] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," Proceedings of the Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, pp. 311-318, 2002.

DUT Verification Through an Efficient and Reusable Environment with Optimum Assertion and Functional Coverage in SystemVerilog

Deepika Ahlawat
VLSI Group
Department of Electrical, Electronics & Communication Engineering, ITM University,
Gurgaon, (Haryana), India

Neeraj Kr. Shukla
VLSI Group
Department of Electrical, Electronics & Communication Engineering, ITM University,
Gurgaon, (Haryana), India

***Abstract*—Verification is the most integral part of chip manufacturing and testing and is as important as the designing. Verification provides with the actual implementation and functionality of a Design under Test (DUT) and checks if it meets the specifications or not. In this paper, a communication protocol has been verified as per the design specifications. The environment so created completely wraps the design under verification and observes an optimum functional and assertion based coverage. The coverage so obtained is 100% assertion based coverage and 83.3% functional coverage using SV (SystemVerilog). The total coverage so obtained is 91.66%.**

Keywords—Assertions; Coverage; Environment; Mailbox; Randomization;SystemVerilog; Threads; Transactions

I. INTRODUCTION

With increasing complexity of the input constraints and the need for better control of the statistical distribution, imperative test benches are being replaced by more declarative specification methods using languages such SystemVerilog [1].

A. Need of Verification

Exponentially increasing complexity of chips particularly SOCs made verification more challenging. Major portion of development time (~70%) of a complex SOC is spent on verification. Reducing verification effort or time spent on verification has a strong impact on Time-to-Market (TTM). In order to satisfy such growing complex verification needs powerful verification languages and verification methodologies are employed [2].

In general IP Verification requires in depth verification with coverage based and constraint random simulation technique, which needs an advanced test bench equipped with various components such as coverage monitors and scoreboards. But if an IP was fully verified before and has a minor design change, it is not necessary to verify all features in detail. A few directed cases and simple checkers might be sufficient [3].

Except for simple cases, the behavioral specification of hardware designs is mostly incomplete, leaving the design's response to many input stimuli undefined. During verification, unspecified inputs must be excluded from examination to avoid undetermined or spurious erroneous behavior. In a simulation-based verification setting, the concept of a "test bench" is applied to specify valid input sequences as well as the expected design responses for them [4].

B. Need of System Verilog

SV is built on top of Verilog 2001. SV improves the productivity, readability, and reusability of Verilog based code. It brings a higher level of abstraction to design and verification. The language enhancements in SV provide more concise hardware descriptions, while still providing an easy route with existing tools into current hardware implementation flows[5].

SV provides a complete verification environment, employing Directed and Constraint Random Generation, Assertion Based Verification and Coverage Driven Verification. These methods improve the verification process dramatically. It also provides enhanced hardware-modeling features, which improve the RTL (Register Transfer Level) design productivity and simplify the design process.

Advantages of Using SV

1) *SV was adopted as a standard by the Accellera organization, and is approval by IEEE. These ensure a wide embracing and support by multiple vendors of EDA (Electronics Design & Automation) tools and verification IP's, as well as interoperability between different tools and vendors [5].*

2) *Since SV is an extension of the popular Verilog language, the adoption process of SV by engineers is extremely easy and straightforward. SV enables engineers to adopt a modular approach for integrating new modules into any existing code. As a result, the risks and costs of adopting a new verification language are reduced.*

3) *Being an integral part of the simulation engine, eliminates the need for external verification tools and interfaces, and thus ensures optimal performance (running at least x2 faster than with any other verification languages) [5].*

4) *SV brings a higher level of abstraction to the Verilog designer. Constructs and commands like Interfaces, new Data types (logic, int), Enumerated types, Arrays, Hardware-specific always (always_ff, always_comb) and others allow modeling of RTL designs easily, and with less coding.*

5) *SV extends the modeling aspects of Verilog by adding a Direct Programming Interface which allows C, C++, SystemC and Verilog code to work together without the overhead of the Verilog PLI (Programmable Logic Interface).*

A declarative description of input constraints is significantly easier to develop in terms of avoiding over constraining or under constraining the inputs as well as controlling the desired distribution. It is expressed as a predicate on the design's input variables such that an input stimulus is valid if and only if the predicate evaluates to true. Advanced test benches must handle cases in which the validity of an input stimulus may differ from design state to design state, which makes the constraints dependent on state variables [4].

The paper is organized as follows, after an overview of verification and advantages of using SV, section II describes the DUT taken and gives a brief introspection on its working. Section III discusses the test bench architecture and all the components it comprises of. Section IV describes how the SV environment works and the various phases of the test bench. Section V consists of the simulation results in the form of waveforms and the coverage report based on assertion coverage and functional coverage.

II. DUT – THE SPI CORE

The serial interface consists of slave select lines, serial clock lines, as well as input and output data lines. All transfers are full duplex transfers of a programmable number of bits per transfer (up to 64 bits). It can drive data to the output data line in respect to the falling (SPI/Microwire compliant) or rising edge of the serial clock, and it can latch data on an input data line on the rising (SPI/Microwire compliant) or falling edge of a serial clock line [6].

Data Transmission

The bus master configures the clock first, using a frequency less than or equal to the maximum frequency the slave device supports. Such frequencies are commonly in the range of 10kHz–100 MHz [6].

During each SPI clock cycle, a full duplex data transmission occurs [7]:

a) the master sends a bit on the MOSI line; the slave reads it from that same line

b) the slave sends a bit on the MISO line; the master reads it from that same line

Transmissions may involve any number of clock cycles. When there is no more data to be transmitted, the master stops toggling its clock. Normally, it then deselects the slave.

Transmissions often consist of 8-bit words, and a master can initiate multiple such transmissions if it wishes/needs. The master must select only one slave at a time [6].

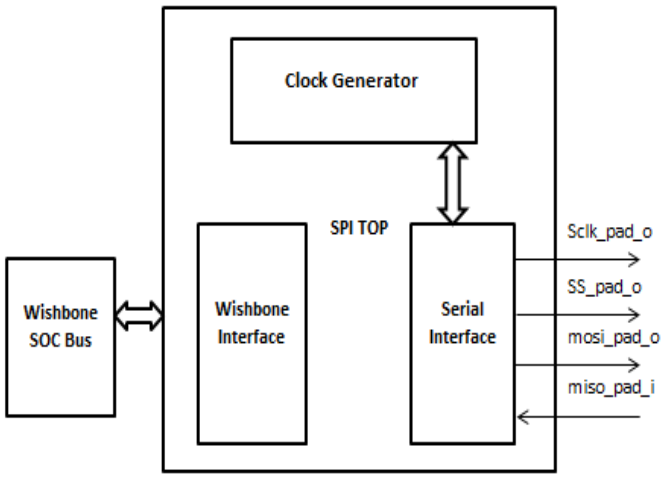


Fig. 1. SPI Architecture [7]

WISHBONE BUS

The Wishbone Busis an open source hardware computer bus, intended to allow parallel communication between the parts of an integrated circuit. This System-on-Chip interconnection architecture is used in order to create a common interface between different IP cores. The Wishbone interconnect is intended as a general purpose interface. As such, it defines a master / slave standard for data exchange between IP core modules, in terms of signals, clock cycles, and high & low levels.

III. SYSTEMVERILOG TESTBENCH ARCHITECTURE

The testbench architecture has various modules as discussed below. The interconnection between these modules can be seen in figure 2.

A. Test Generation

A Test case is a program block which provides an entry point for the test. The test case generator will provide all the valid test cases to the driver. The test cases are generated by randomizing certain inputs and registers while keeping some fixed.

To perform this type of randomization i.e. constraint randomization a function called random is created [8].

B. Driver

The driver will reset and configure the DUT.

It will call the tasks from test generator and will form a packet in the packet generator module and will unpack the packet in the driver module and implement it on the DUT. The interfaces of the Driver are: clk_i, rst_i, add_i, data_i, sel_i, we_i, stb_i, cyc_i, sclk, miso.

C. Monitor

Monitor will keep track of all the test cases provided to the driver. It will also look at all the signals coming from the DUT. Monitor thus will call the packet in the scoreboard module and compare it with the output from the DUT present in the checker. Hence the duty of monitor is to complete simulation when all cases have been read. It will also generate error message if there is any discrepancy in data out coming from DUT and the teat generator. The interfaces of the monitor are: data_out,int_o, ack_o, ss, err_o, mosi.

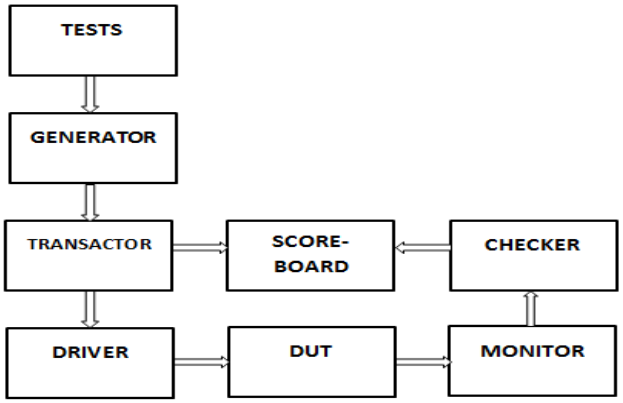


Fig. 2. Verification Flow [8]

D. Responder

Responder is a block which acts as a slave and gives out miso_pad_i to the DUT which is processed or stored or read from the DUT. It is given sclk or slave clock from the DUT and sends miso to the core.

E. Scoreboard

The output from monitor is checked with the expected output. Ihe output genetared by the DUT as observed by the monitor is passed to the scoreboard through mailbox. If the actual output does not match the expected output an error message is generated else if it matches a pass message is displayed.

F. Coverage

Coverage will check the functional coverage of the DUT by the test cases tested by the driver and monitored by the monitor. It will also create an error counter which will show the TEST FAIL and TEST PASS status [8].

IV. Compilation in SystemVerilog

Following are the methods which defined in the environment class of the SV testbench[6].

A. build (): In this method , all the objects like driver, output monitor and mailboxes are constructed.

B. reset (): in this method all the signals are put at a known state.

C. start (): in this method, all the methods which are declared in the other components like driver, output monitor and scoreboard are called.

D. wait_for_end (): this method is used to wait for the end of the simulation. Wait is done until all the required operations in other components are done.

E. report (): This method is used to print the results of the simulation, based on the error count.

F. run (): This method calls all the above declared methods in a sequence.

The way the DUT interface with the driver, monitor and responder/slave can be seen in figure 3.

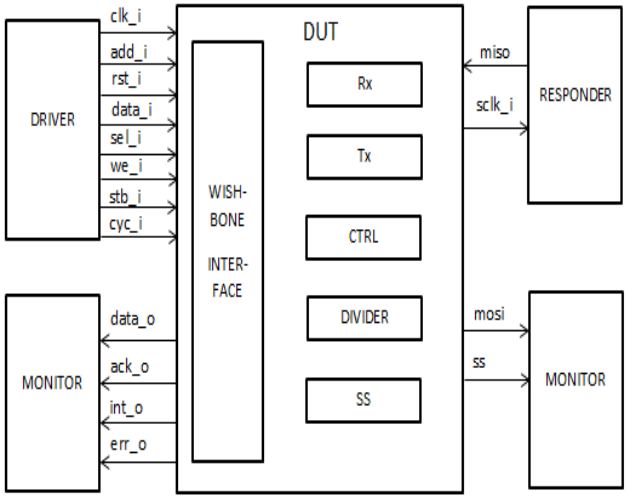


Fig. 3. Architectural overview of the verification modules as implemented in the proposed verification environment

V. Design Simulation

A. Randomization

Random testing is more effective than a traditional approach of directed testing. One can easily create tests that can find hard-to-reach corner cases, by specifying constraints. SystemVerilog allows users to specify constraints in a more compact and declarative way. The constraints are then processed by a solver that generates random values that meet the constraints [5]. The stimuli randomizes are data input, slave select and address input as seen in figure 4.

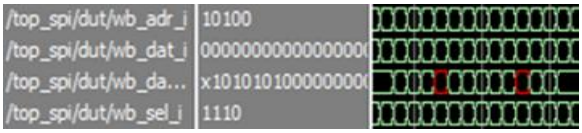


Fig. 4. Randomized value of signals

B. DUT Signals Generated

ack

The acknowledge output [ack_o] indicates the normal termination of a valid bus cycle.
The ack signal obtained can be seen in the figure 5 below.

Fig. 5. Acknowledgment signal generated

Sclk

SCK [sck_o] (figure 6) is generated by the master device and synchronizes data movement in and out of the device through the MOSI [mosi_o] and MISO [miso_o] lines. The SPI clock is generated by dividing the WISHBONE clock [clk_i].

Miso

The Master In Slave Out line is a unidirectional serial data signal. It is an output from a slave device and an input to a master device (figure 6).

Mosi

The Master Out Slave In line is a unidirectional serial data signal. It is an output from a master device and an input to a slave device (figure 6).

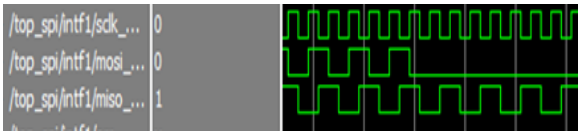


Fig. 6. sclk, miso and mosi signals generated

C. Output Waveform

The output waveform as shown in figure 7, displays all the signals being generated by the DUT. The internal registers are also seen to be crunching data and displaying corresponding outputs through SPI signals.

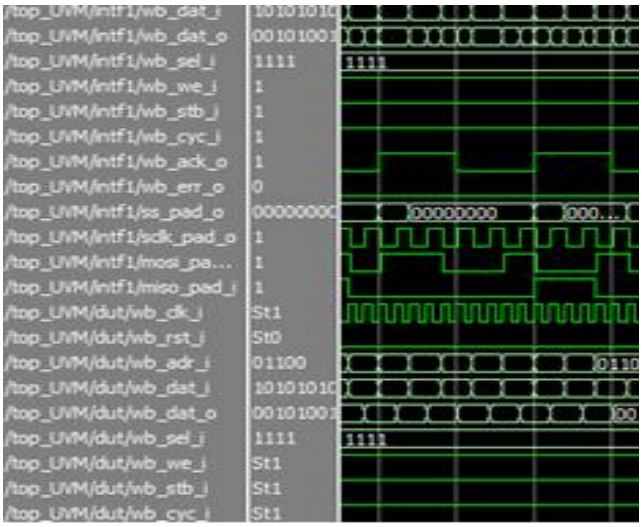


Fig. 7. Output Waveform of SPI Core

D. Coverage

1) Assertion Coverage

Assertions are mechanism or tool used by HDL's (VHDL and Verilog) to detect a design's expected behavior. The assertion fails if a property that is being checked for in a simulation does not behave the way we expect it to or we can say that, the assertion fails if a property that is forbidden from happening in a design happens during simulation.

It helps capturing the designer's interpretation of the specification [5]. The assertion coverage based on randomization function assertion is shown in figure 8.

Coverage Summary by Structure:

Design Scope	Coverage (%)
testcase_sv_unit	100.00%
generator	100.00%

Coverage Summary by Type:

Weighted Average:				100.00%
Coverage Type	Bins	Hits	Misses	Coverage (%)
Assertion Attempted	1	1	0	100.00%
Assertion Failures	1	0	-	0.00%
Assertion Successes	1	1	0	100.00%

Fig. 8. Coverage report based on assertion

2) Total Coverage Percentage

Total coverage here (figure 9) includes both the assertion based coverage and the functional coverage. The functional coverage is based on the coverpoints of the corresponding covergroup. Bins have been created and have been hit properly to generate functional coverage.

Coverage Summary by Structure:

Design Scope	Coverage (%)
testcase_sv_unit	91.66%
generator	91.66%

Coverage Summary by Type:

Weighted Average:				91.66%
Coverage Type	Bins	Hits	Misses	Coverage (%)
Covergroup	6	5	1	83.33%
Assertion Attempted	1	1	0	100.00%
Assertion Failures	1	0	-	0.00%
Assertion Successes	1	1	0	100.00%

Report generated by Questa on Saturday 16 November 2013 21:09:00

Fig. 9. Coverage report including functional and assertion based coverage

VI. Conclusions

The code for environment has been simulated. The outputs from DUT have been observed. Environment contains the instances or the objects of the driver, monitor, scoreboard and the DUT. The task performed by the monitor, driver and scoreboard is called along with the mailboxes which contain the received and sent information in the form of randomized packets. The mailbox implemented to carry the packets shows results after every transaction. The environment so created completely wraps the design under verification and observes an optimum functional and assertion based coverage. Bins have been created based on the constraints and 85-100% functional coverage has been obtained on them. The coverage so obtained is 100% assertion based coverage and 83.3% functional coverage using System Verilog. The total coverage so obtained is 91.66%.

Acknowledgment

The authors are grateful to their respective organization for help and support.

References

[1] Sutherland S, Davidmann S, Flake P, "SystemVerilog for Design: A Guide to Using SystemVerilog for Hardware Design and Modeling," Norwell, MA: Kluwer Academic Publishers, 2003.

[2] SudhishNaveen, BR Raghavendra, YagainHarish, "An Efficient Method for Using Transaction Level Assertions in a Class Based Verification Environment," International Symposium on Electronic System Design,pp.72-76, 2011

[3] Yun Young-Nam, Kim Jae-Beom, Kim Nam-Do, Min Byeong, "Beyond UVM for practical SoC verification," SoC Design Conference (ISOCC), pp. 158 – 162, Nov 2011

[4] Welp Tobias, Kitchen Nathan, and Kuehlmann Andreas, "Hardware Acceleration for Constraint Solving for Random Simulation,"IEEE Transactions On Computer-Aided Design of Integrated Circuits And Systems, vol-31, No. 5, May 2012

[5] [Online]Available: http://www.systemverilog.in/systemverilog_introduction.php

[6] K.Aditya,M.Sivakumar,FazalNoorbasha, T.PraveenBlessington, "Design and Functional Verification of A SPI Master Slave Core Using System Verilog," International Journal of Soft Computing and Engineering (IJSCE), vol-2, Issue-2, May 2012

[7] SrotSimon, "SPI Master Core Specification,"Rev. 0.6, March 15, 2004

[8] RaoAbhiram. *What is SystemVerilog?*[Online] Available:http://electrosofts.com/systemverilog/introduction.html

ABOUT THE AUTHORS

DeepikaAhlawat,completed her B.Tech in Electronics and Communication Engineering from Gurgaon College of Engineering for Women, Gurgaon in 2012. She is now pursuing her Master of Technology (M.Tech) in VLSI Design at ITM University, Gurgaon. Her interest includes Digital Design, ASIC Design, VLSI Testing and FPGA prototyping.

Dr. Neeraj Kr. Shukla (IEEE, IACSIT, IAENG, IETE, IE, CSI, ISTE, VSI-India), an Asst. Professor in the Department of Electrical, Electronics & Communication Engineering, ITM University, Gurgaon, (Haryana) India. He has received his M.Tech. Degree in Electronics Engineering and B.Tech. Degree in Electronics & Telecommunication Engineering from the J.K. Institute of Applied Physics & Technology, University of Allahabad, Allahabad (Uttar Pradesh) India in the year of 1998 and 2000, respectively. His main research interests are in Low-Power Digital VLSI Design and its Multimedia Applications, Digital Hardware Design, Open Source EDA, Scripting and their role in VLSI Design, and RTL Design.

Towards a Modular Recommender System for Research Papers written in Albanian

Klesti Hoxha, Alda Kika, Eriglen Gani, Silvana Greca
Department of Computer Science
University of Tirana, Faculty of Natural Sciences
Tirana, Albania

***Abstract*—In the recent years there has been an increase in scientific papers publications in Albania and its neighboring countries that have large communities of Albanian speaking researchers. Many of these papers are written in Albanian. It is a very time consuming task to find papers related to the researchers' work, because there is no concrete system that facilitates this process. In this paper we present the design of a modular intelligent search system for articles written in Albanian. The main part of it is the recommender module that facilitates searching by providing relevant articles to the users (in comparison with a given one). We used a cosine similarity based heuristics that differentiates the importance of term frequencies based on their location in the article. We did not notice big differences on the recommendation results when using different combinations of the importance factors of the keywords, title, abstract and body. We got similar results when using only the title and abstract in comparison with the other combinations. Because we got fairly good results in this initial approach, we believe that similar recommender systems for documents written in Albanian can be built also in contexts not related to scientific publishing.**

Keywords—recommender system; Albanian; information retrieval; intelligent search; digital library

I. INTRODUCTION

Relevant information retrieval is very important for the scientific community, but also a very time consuming task. The academic search engines usually use keywords to find the relevant articles. This approach often produces unsatisfying results. An alternative approach suggests the usage of a recommender system to facilitate the retrieval of relevant information [1] to potential users. A recommender system assists the users in the process of finding relevant and personalized information fast.

Recommender systems are designed to help users navigate through complex and overload information by suggesting which items a user could have interest [1]. They are used in many domains, including music, movies, TV programs, videos on demand, books, news, images, web pages, research papers etc. Their role in our information society is becoming essential. The interest in this area is high because recommender systems help in the process of localization of personalized information from the overload gathered data.

In Albania there are many scientific journals and conference proceedings, which produce a lot of scientific papers in Albanian language. The scientific papers are published in hard copy journals, optical media (i.e. CD-ROM) and in the corresponding journal web pages. The Albanian researchers usually search the articles related to their work through traditional digital means using specific keywords or even hand browsing the individual websites of particular journals or conferences. This labor-intensive task in searching for articles is often useless and the retrieved articles may not be the needed ones. The search task in the hard copy journals is even more exhaustive. Recommender systems can provide a considerate help to Albanian researchers to acquire proper information from digital libraries.

In this paper we propose the design of a system that considers the case of a scientific journal which has begun as a hard copy journal and now is in the process of creating a web application in order to publish and access the published papers which can be in the English and Albanian language. The fact that most of the scientific articles are written in Albanian makes the process more difficult since there are no concrete systems which deal with information retrieval tasks regarding documents written in Albanian. The proposed recommender system is designed in a modular form, enabling easy replacement and modification of the components and experimenting with new algorithms in the future.

To detect the similarity between the interest of users and the available resources different approaches can be used. Most of them make use of the vector space model [16] that represents the items in question as vectors in a vector space. Each dimension of this vector space represents a feature of the items. When calculating the similarity of items, a possibility is to calculate the distance between these vectors for example by using the cosine similarity described in [4]. When recommending items to users, they are usually ranked in decreasing similarity order in comparison to other vectors in the items vector space. The different recommendation approaches differ the way this vector (or a set of vectors) is chosen [1]. It may be an item that is known beforehand that is in the users interests, or it might consist of a set of preferences of the user stored as a vector in the same vector space of the items. Another suggested approach is to recommend to a particular user items liked by users with similar profiles to him [6].

The focus of this paper is to present the design of a system with a highly modular architecture and make the first steps towards a recommender system that recommends documents (scientific articles in our case) written in Albanian. We used a periodic scientific journal published in Albania as a testing dataset. It contains scientific papers written mostly in Albanian

about five different research areas: mathematics, physics, biology, chemistry, and computer science. We did not have to crawl the web for collecting these articles because we had access to the articles' collection. For building the recommender system we made use of the cosine similarity measure which is generally used in various recommender systems for digital libraries [2, 3, 23]. The used dataset was relatively small; it consisted of 226 articles in total. However the aim of this work was to identify the experimentation settings that produce the better results with the chosen similarity measure.

Dealing with papers written in Albanian makes necessary the usage of a word stemmer [16] designed for this language. We used the Albanian stemmer suggested in [17] that has been successfully tested in a document classification context. In our recommender system we used this algorithm in two different ways: stemming words with a single run and stemming with multiple runs over the initial word. This experiment was performed because some Albanian words can be further reduced after the first run of the stemming algorithm proposed in [17].

We gained fairly good results in our experiments. There were relevant items in the list of articles recommended to the users. In the performed experiments, we differentiated the importance of each part of the article in the used similarity heuristics. Our results showed that there were no big differences among the produced results. It was demonstrated that the experiments that used only the abstract and title of the articles for the recommendation process produced better or same results as most of the other performed experiments. This result was gained also by Nascimento et al. [23]. In terms of computation this speeds up the recommendation process, because the processed text (title and abstract) consists of a small part of the article.

The achieved results produced positive insight about future improvements of the system, or the creation of new information retrieval systems that deal with nonscientific related documents (i.e. news articles, laws) written in Albanian.

The first part of this paper briefly presents some of the research literature related to the existing approaches of designing the recommender system. The other parts introduce the proposed system architecture, the technologies that have been used, and the similarity heuristics that have been tested. The paper is concluded by presenting the conclusions and future work.

II. RELATED WORK

The three basic approaches used in the design of recommendation systems are: content-based, collaborative filtering and hybrid [1]. The content-based recommender systems [5] default strategy consists of matching up the previously collected attributes of a user profile with those of the items in question, with the intent to arrive at a relevant result. This comparison is usually done in a vector space that stores the items and the user profiles as feature vectors [16]. Each dimension of this vector space represents a particular item feature.

Another possible content-based recommendation system strategy analyzes item descriptions to identify items that are of particular interest to the users. The filtering techniques used in this approach rely on item descriptions and generate recommendations from items that are similar to those that the target user has liked in the past, without directly relying on the preferences of the users (stored in their profiles) or other individuals [5]. This last approach does not require a large user base and collected data about them. When lacking the latest, the collaborative-filtering approach (see below) would be ineffective. The content-based approach makes possible a recommendation based solely on the description of the items themselves and not the interested users. This is the case for the initial stage of most digital libraries and similar information retrieval systems [23]. The similarity comparison in this approach is straightforward because it compares an item with other items, however in order for the recommendation to make sense, there is the need for an initial item (article) that is known to be of some interest for the user.

Collaborative filtering recommender systems [6] recommend items based on the past preferences of similar users. The recommendation is based on the assumption that items liked by users with similar profiles to a concrete user, are highly probable to be liked by the latest. This requires having a solid user profile database that stores the preferences and activity related data about the users. If the users of the digital library are not actively participating by making reviews or providing some feedback about the articles, or if they do not have full specified profiles (research area, interests), this database would lack of important data for the recommendation process. However if these data exists, there is a high probability that the recommendation process produces good results [6], [7].

Hybrid recommender systems [7] usually use a combination of content based and collaborative filtering recommendation for recommending items. This combined approach deals with the drawbacks of the above described ones, allowing for an initial content-based recommendation in cases of a cold start (lack of user profiles) [23]. The collaborative-filtering recommendation can improve the results by adding context-related information to the content-based approach.

Although recommender systems are very popular in commercial applications these days, recommender systems for the academic research have also gained interest. This is noticed by the emergence of a lot of research papers about this topic presented in many conferences and journals. Below we describe some of the applications of recommender systems in scientific paper recommendation situations.

Docear is an academic literature suite to search, organize, and create research articles [8]. Its recommender system [9] uses content based filtering methods to recommend articles. It allows the users to build "mind maps" that represent a user model (profile) which is matched with Docear's Digital Library. The authors claim to have achieved decent results based on the number of clicks gained through about 30 thousand tested recommendation results.

In [10] a personalized academic research paper recommendation system is presented. It recommends relevant articles to the research field of the users. It is supposed that the users (researchers) "like" their own articles. Based on this

assumption papers similar to the ones previously written by the system users are recommended as relevant to them. This system uses a web crawler to retrieve research papers from two concrete digital libraries: IEEE Xplore and ACM Digital Library. It uses text similarity to determine the similarity between two research papers and collaborative filtering methods to recommend the items.

Nascimento et al. provide another example of a content-based recommender system for scientific articles [23]. They point out that most of the recommender system approaches suppose that a large collection of scientific papers is available beforehand. This is the case for some digital libraries like IEEE Xplore, but it does not hold for many other situations. Their proposed solution depends on publicly available scientific metadata, concretely the title and abstract of the articles. Their designed system collects these data by simulating searches on the websites of various publishers. Instead of using user defined keywords, they generate keywords from a particular article that is presented by the users (most probably an article in their particular research area).

The similarity of the articles is calculated by using the cosine similarity based on the vector space model [4]. The same similarity measure is used in our designed recommender system (see below). The results gained by Nascimento et al. were fairly positive, demonstrating that it is enough to consider only the title and abstract of the articles for recommendation purposes.

The hybrid approach of recommender systems has also been used for recommending research papers [11, 12]. As an example we have Techlend, in which different techniques of combining content-based and collaborative-filtering recommendation algorithms have been compared [11]. The experiments were sequential ones. The results of one algorithm were fed to the other. In general this approach produced good results. The test dataset was quite large, including about 100 thousand research papers indexed in CiteSeerX[1]. In 85% of the cases, the users found at least one related article to the presented recommendation list. Because some of the performed experiments did not perform well, the authors suggest that the sequential execution of the two involved recommendation algorithms (content-based and collaborative-filtering) is not the best alternative.

Another approach used by some academic paper recommender systems uses the paper's citations for recommending articles. In [12] it is presented another hybrid recommendation system. It aimed to be a powerful alternative to academic search engines by not solely relying on keyword analysis, but by additionally using citation analysis, explicit ratings, implicit ratings, author analysis, and source analysis. The popular academic search engine CiteSeerX also uses citations to find similar scientific papers [26]. Some other applications with citation recommendation are presented in [13], [14], and [15].

To the best of our knowledge, there have been no serious works regarding intelligent recommender systems that deal with documents written in Albanian. Maybe the main reason was the lack of enabler tools written specifically for the Albanian language that are used by many information retrieval systems, i.e. word stemmers and part of speech taggers [16]. However, the situation seems to have changed recently. A few Albanian language stemming algorithms have emerged, for example the ones described in [17] and [18]. The same holds for part of speech (POS) taggers, as examples we have the ones proposed in [19] and [20].

The stemming algorithm proposed by Sadiku and Biba in [17] has been tested in classifying documents written in Albanian about biology, history, literary and chemistry. They noticed an accuracy increase when using the stemming algorithm in comparison when it was not used. They also pointed out that the results were worsened when classifying documents of related fields. This fact is not strictly connected to the stemming algorithm itself, biology and chemistry articles have similar words in their content.

III. Context Description

In this paper we provide the design of an intelligent search system about academic papers written in Albanian. Right now it is very difficult for Albanian researchers to find works related to their field of research, present in several journals or conference proceedings published in Albania and its neighboring countries that have large communities of Albanian speaking researchers.

We aim to provide a system that not only allows for keyword based searching of scientific papers, but also recommends related articles (to a concrete article). In this case we suppose that the user "liked" a certain article, after finding it by a normal search (i.e. via keywords) and is interested in finding other articles similar to this one. The first requirement of our system is the creation of an index of articles that stores metadata about each of them and enables searching. The created index also allows for further investigations about a collection of articles written in Albanian, like topic identification [6] and research trends detection.

Our testing dataset was obtained from a periodical scientific journal published by the Faculty of Natural Science, University of Tirana, Albania[2]. It contains scientific articles about five main research fields: mathematics, physics, biology, chemistry, and computer science. Most of these articles are written in Albanian. They are provided in PDF format and follow some standard formatting guidelines (font weight, font size, etc.).

We aimed on increasing the visibility of each individual article of the above mentioned journal, allowing researchers to easily find scientific articles related to their research work. Also, because scientific articles have all a very similar structure, our implemented system can be used for other articles written in Albanian. This addresses a crucial need of the Albanian research community and would boost the quality of research work done in Albania and its neighboring countries.

[1] http://citeseerx.ist.psu.edu/

[2] The Bulletin of Natural Sciences (Buletini i Shkencave Natyrore), http://buletini.fshn.edu.al/

IV. Proposed System Architecture

We designed a highly modular system architecture that allows for louse coupling of the proposed modules (shown in Figure 1). Each individual module is designed to be as much independent as possible, so we can easily extend or change the behavior of the actual system in the future. It is to a high degree independent of the actual scientific article formatting (it is possible to import multiple types of article layouts after a correct specification of the formatting rules).

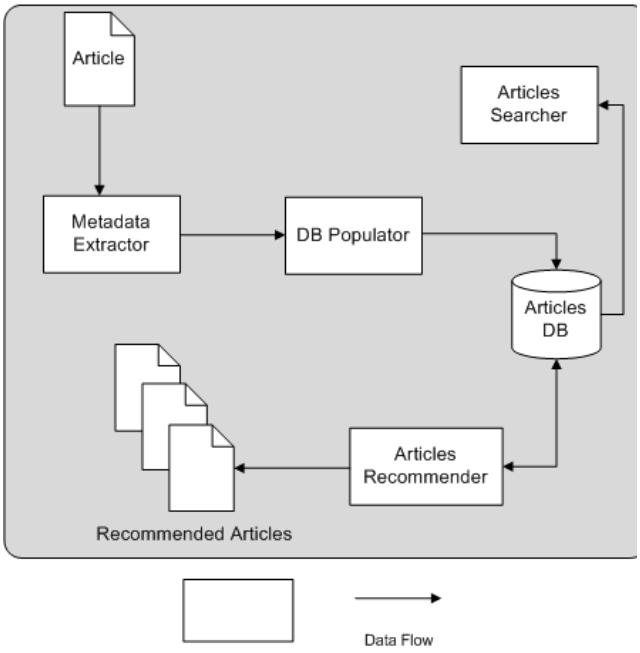


Fig. 1. Proposed system architecture

We have build our system by using Java EE 7, taking advantages of Java Persistence API (JPA) for reading and manipulating the articles metadata stored in the articles database. Of course, the initial step for using our system consists of uploading the actual articles to it. We uploaded the articles of the Bulletin of Natural Sciences in PDF format. However, there is no actual limitation from our system regarding the file type of the articles. It should be possible to support all the major file types, of course by providing an appropriate parser.

The *articles database* stores metadata about the articles, like the title, abstract, authors, keywords, body, and the article's parts. It also stores the location of the article files (PDF for the Bulletin of Natural Sciences) in the server, so that they can be provided via a user interface to the system's users. An important metadata that we store is the term frequency for each individual term, i.e. the number of times an individual term (word) appears in a single document [16]. We store the term frequency related to the actual article parts. We differentiate between "body term frequency", "title term frequency", and "abstract term frequency", and "article parts term frequency" respectively counting the frequency of a term within the body, title, abstract, and each individual part (section) of an article.

The term frequencies stored in the database are not weighted, but a term weighting scheme [16] can be derived easily by the stored frequencies, as will be shown below. We used a normalized relational database schema for storing the metadata, implemented in MySQL (using its InnoDB engine).

The *database populator* is the part of the system that stores the extracted metadata in the database. It uses the *metadata extractor* module, but it is not dependent on the actual article parser. We used JPA as an abstraction layer responsible for storing the data in the database.

The *metadata extractor* is responsible for extracting the metadata by the article files. Several parsers can be implemented according to the formatting of the articles. This module does not use any machine learning approach for detecting the article components (title, abstract, etc.) automatically from any kind of article like in [24]. We use a simpler approach, because the scope of this article is not to provide a general metadata parser from any kind of article format. In our approach for the Bulletin of Natural Sciences, we parse the PDF files of each individual article based on the formatting guidelines (text size, font weight, etc.) of them. We use *pdfbox*[3] as a PDF parser library and we extract the metadata directly from the PDF (without converting it to any other format, like text or xml). The metadata extractor makes use of an Albanian language stemmer, i.e. a software component that reduces a given word in its "stem", the part of the word that does not contain any suffixes or postfixes [16]. For example the Albanian word *bashkëpunoj* (collaborate) is reduced to *pun(ë)* (work). We used the algorithm described in [18] for stemming the words (terms). Furthermore we removed a list of stop words, words that appear most frequently in the Albanian language, based on a combined list of stop words provided by [18] and [22]. This step is critical, because the most frequent words of the Albanian language would affect in a large degree the heuristics described below that we used for finding similar articles. It is easy to change the stemming algorithm in use by the module if it is needed, this also holds for other tools used in natural language processing (NLP), like part of speech taggers described in [19] and [20].

The *articles searcher* is used when searching articles by using keyword based queries. It uses the metadata stored in the database as an index and returns search results based simply on the presence or not of a term in a particular document. The results are ranked by the frequency of the searched term (terms) in the document. The complete description of the articles searcher functionality is out of the scope of this article.

The *articles recommender* is the part of our system that behaves like a recommender system [1]. It recommends similar articles to the one that the user is currently viewing. The aim is to facilitate the discovering of articles that are about similar research questions. The similarity of articles is calculated by using a heuristics that considers the provided keywords, term frequencies located in the title, the abstract, and the body of the articles. We have implemented a content-based recommender system [5] because we lack of user profiles at this stage, and

[3] http://pdfbox.apache.org/

most importantly user activity information. The complete similarity calculation details are provided in the next section. This module makes use only of the articles metadata database and it is independent of the metadata extractor itself.

For the articles recommender, metadata extractor, and articles searcher module we have also provided appropriate web services that allow for easy integration of our system with third-party articles publishing systems (i.e. other scientific journals).

The articles can be fed into the system by using the provided web service. We do not follow a web crawling approach like in [10] and [23]. The recommendation is done offline, based on the collected dataset.

V. Similarity Calculation Heuristics

For the articles recommender module (see Figure 1) we have implemented some heuristics that calculate the similarity of two articles based on the keywords provided by the authors and the term frequencies on their respective title, abstract, and body. More concretely, in (1) we show the metrics we have used for similarity calculation. It consists of the cosine similarity of the vector space model [16], used successfully for scientific articles comparison also in [2], [10] and [23]. The similarity is calculated considering that each document is represented as a vector within a vector space whose dimensions consist of the weighted term frequencies (for each term).

$$sim(d_1, d_2) = \frac{\vec{V}(d_1)\cdot\vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|} = \frac{\sum_1^k w_{1i} w_{2i}}{\sqrt{\sum_1^k {w_{1i}}^2}\times\sqrt{\sum_1^k {w_{2i}}^2}} \quad (1)$$

w_{ji} consists of the weighted term frequency of term i in document d_j. To calculate the weighted term frequencies we use the heuristic presented in (2). It uses the raw term frequencies stored in the articles metadata database. As explained in section IV, the terms have been stemmed beforehand and stop words have been removed. The term frequencies of term i in the keywords list, title, abstract and body of article j, are denoted respectively as w^k_{ji}, w^t_{ji}, w^a_{ji}, w^b_{ji}. We used the "terms keywords list frequency" based on the assumption that keywords chosen by the authors of an article, are the most representative terms of the content of it. w^k_{ji} equals 1 if term i is present in document j keywords list. The title, abstract and body term frequencies have been weighted using the tf-idf scheme [15], lowering the "importance" of terms that appear too often in the whole articles' collection.

$$w_{ji} = \kappa\, w^k_{ji} + \tau\, w^t_{ji} + \alpha\, w^a_{ji} + \beta\, w^b_{ji} \quad (2)$$

The coefficients κ, τ, α and β are set according to the importance of each article part (keywords, title, abstract, body) in the term frequency calculation. Because w_{ij} is an affine linear combination, then $\kappa + \tau + \alpha + \beta = 1$.

Given a single article, the system calculates the similarity of it with each of the other articles of the collection. Then the results are sorted in descending similarity value order and the top x similar articles are showed to the user (i.e. top 10). We generate the recommended articles by using a background job that stores the results in the articles metadata database. Due to our highly modular system architecture, it is possible to easily change the similarity function that is used to any possible similarity measure. Our testing dataset, the Bulletin of Natural Sciences, contains articles from different research categories (mathematics, computer science, biology, etc.), therefore we limited the similar document search within articles of the same category (i.e. biology).

VI. Experiments and Evaluation

In order to test our system we ran some experiments that altered the coefficients used in (2) and the way the stemming algorithm works. The way we chose the used coefficients tried to diversify the importance of each article part based in common sense and experimentation purposes. Concretely we performed three base experiments with the following importance coefficients:

1) $\kappa = 0.4, \tau = 0.3, \alpha = 0.2, \beta = 0.1$, giving more importance to the keywords list and title terms of an article

2) $\kappa = 0.0, \tau = 0.6, \alpha = 0.4, \beta = 0.0$, excluding the keywords and using only the title and abstract term frequencies

3) $\kappa = 0.4, \tau = 0.0, \alpha = 0.0, \beta = 0.6$, using only the keywords and body term frequencies for similarity calculation

Furthermore we used two different stemming strategies using the algorithm proposed in [17]:

1) We stemmed the words by using a single run (pass) of the stemming function.

2) We stemmed the words by using several iterations of the stemming function, until the word cannot be stemmed anymore. The ways the Albanian words are constructed creates cases that a word can be stemmed again after the first run of the stemming algorithm in use.

It should be noted that our testing dataset contains only 226 scientific articles written in Albanian with the following distribution: 19 articles belong to the physics category, 22 to the mathematics category, 25 to the computer science category, 78 to the chemistry category, and 82 to the biology category.

Our experiments did not aim to measure the performance of the used strategy in terms of execution time, but only the relevance of the recommended articles to the input article.

We used standard evaluation measures of information retrieval systems: *precision*, *recall* and F_1 (a combination of precision and recall) [16], defined as in (3), (4), and (5).

$$Precision = \frac{\#(relevant\ items\ retrieved)}{\#(retrieved\ items)} = P \quad (3)$$

$$Recall = \frac{\#(relevant\ items\ retrieved)}{\#(total\ number\ of\ relevant\ items)} = R \quad (4)$$

$$F_1 = (2PR)/(P + R) \quad (5)$$

We calculated the evaluation measures for each combination of the parameters described above in this section. The experiments consisted of reviewing the recommended articles of 10 random articles of the collection. In order to find

the number of relevant items, each of the recommendations was rated as "related" or "not related". This evaluation scheme is similar to the one used in [23], however we did not introduce more than one level of relevance (slightly, very, etc.). The total relevant recommendations number required in (4) was calculated by considering the total number of relevant articles found in all of the performed experiments.

TABLE I. EXPERIMENTS RESULTS

Stemming	**Coefficients used**		
	$\kappa = 0.4$ $\tau = 0.3$ $\alpha = 0.2$ $\beta = 0.1$	$\kappa = 0.0$ $\tau = 0.6$ $\alpha = 0.4$ $\beta = 0.0$	$\kappa = 0.4$ $\tau = 0.0$ $\alpha = 0.0$ $\beta = 0.6$
Single run stemming	P = 0.31 R = 0.18 F_1 = 0.23	P = 0.34 R = 0.20 F_1 = 0.25	P = 0.32 R = 0.18 F_1 = 0.23
Multiple run stemming	P = 0.26 R = 0.15 F_1 = 0.19	P = 0.29 R = 0.17 F_1 = 0.21	P = 0.21 R = 0.12 F_1 = 0.15

The results of the experiments are displayed in Table 1. The first thing that can be noticed from the results is the fact that in general, the single run of the stemming function performs better than the multiple run. This might have happened because some words in the Albanian language may lose their real meaning when stemmed consecutively (like in our second approach).

Regarding the coefficients used in the similarity heuristics given in (2), no big differences are noticed within the experiments performed with a single run stemming. For the experiments performed with the multiple run stemming, the experiment that used only the body and keywords for calculating the weighted term frequencies, resulted the worst performing.

An interesting outcome is the fact that the weighting scheme that used only the title and abstract, performed slightly better than the ones that used also the keywords. Even though we assumed that the keywords chosen by the authors may be the most representative terms of an article, it seems that no real advantage is gained by using them for similarity calculation. This might be an indication that manually chosen keywords do not help very much a recommender system, even though they might produce good results in document classification or automated sorting.

We did not get better results by using the term frequencies of the body of the articles. This fact can be used for improving our recommender system by reducing the size of the index and also the computation time needed for calculating the term frequencies of the body, or making use of them during the similarity calculation.

Nascimento et al. [23], achieved similar results when using only the title and abstract for recommendation calculation. They also did not notice improvements when considering the term frequencies of the body of the articles.

Even though our offline calculation of the recommendations made easier the calculation of the body term frequencies in comparison with the Nascimento et al. approach, we did not gain direct benefits from it.

The evaluation scheme that we used, does not take into consideration the order of the recommended articles. Actually, when considering related items, maybe the most important factor is that the top x results contain the most relevant items. We did get fairly good results. In most of our experiments, there were many related articles in the recommendation list presented. This is an indication that the used heuristics combined with the stemming algorithm presented in [17] and the stop word lists provided in [17] and [22] can produce good results in a recommender system context. So the results showed that the stemming algorithm that was tested in a document classification context in [17], also produced good results in other information retrieval applications.

During the design of our system we also noticed some other terms that can be used as stop words, or possible improvements to the stemming algorithm. However, the tf-idf weighting of the terms reduced the importance of common words in our data set (i.e. the Albanian word sistem = system in English). It is out of the scope of this work to provide a better stemming algorithm for the Albanian language, but we believe that some custom tweaks to it that consider the most used words in a scientific domain might have produced better recommendation results.

VII. CONCLUSIONS

Many recommender systems that help researchers on finding scientific articles related to their work have been recently proposed. There have been different approaches that usually go into the same line as recommender systems used in other areas like e-commerce, movie databases, etc. [1]. Due to the lack of the needed resources for decent information retrieval systems, there have been not much works that deal with documents written in Albanian, and even less in the scientific papers recommendation domain.

In this paper we proposed the design of a highly modular system that indexes and allows for searching of scientific articles written in Albanian. Its modular architecture simplifies the extension of it in the future, and the web services offered allow for easy integration with third-party information systems (i.e. digital libraries).

A crucial part of our designed system is the *articles recommender module*. It is a typical recommender system that recommends to a user a list of related articles (about a given single article). This facilitates a lot article searching, because after finding a particular one, other articles related to it are displayed.

In our approach we built a content-based recommender system [5] that uses cosine similarity of the vector space model for similarity calculation. Because of the similar structure of scientific articles (title, abstract, body, keywords) we used a weighting heuristics that is made of a linear combination of the term frequencies of the title, body and abstract of an article. We also used the keywords list in this heuristics, based on the

assumption that keywords provided by the authors of an article, are the most descriptive terms of it.

The evaluation of our system tried different combinations of the importance factors by using a heuristic described in (2). Our goal was to identify the setting that produces better results. We also tried two different approaches of the stemming algorithm described in [17], single and multiple run of the stemming function on a given word. Our results showed that using a single run of the stemming function produced better results. Also no big differences were noticed within the results gained by the different combinations of the importance factors (coefficients) used. Nevertheless, it was showed that the body term frequencies can be excluded from the heuristics and the results will not change. This was a confirmation of the facts presented in [23].

The usage of the keywords list in the similarity heuristics did not perform better than the case that used only the title and abstract term frequencies. This fact might be an indicator that keywords defined by the authors of the article do not help very much in recommender systems scenarios.

In general, our system performed fairly well, providing relevant articles in each experiment that we made. Even though the used dataset was the one of a particular journal published in Albania, the similar structure of scientific articles allows for usage of other scientific articles, but a custom text parser need to be provided. Also, because the stemming [17] and stop word removal produced good results, we believe that similar information retrieval systems can be built in contexts not related to scientific publishing.

VIII. FUTURE WORK

The lack of user profiles at this stage stopped us from trying a collaborative-filtering approach for our recommender system. We plan to extend the system in the future, introducing user profiles and user feedback. This way we can further improve the search experience and make another step towards an intelligent academic paper recommender system for articles written in Albanian.

We noticed possible improvements to the recommendation results if some further words have been excluded from the calculations, or have been better stemmed. We plan to test a part of speech tagger for Albanian, in order to exclude verbs in the vector space model of the articles. We believe that this might further improve the results.

The system that we built requires that articles should be uploaded to the system beforehand. A different approach would be to index articles found at the websites of other journals, and use our system only as a search engine that links to them.

Another approach that might be tried in the future is the testing of other similarity functions used in the information retrieval domain [16].

The index we created allows for further investigation of the articles dataset. It might be interesting if we can get other information from it, like research trends of the authors and topic identification [21]. Another interesting approach would be to include the authors' research trends in the recommendation formula, i.e. include articles in the recommendation list written by authors that have written articles in the topic of the article in question.

Lastly, the small dataset that we used for testing the system did not allow for careful evaluation of the performance (in terms of execution time) of the system. This will get possible when a larger dataset of articles written in Albanian will be available. At that stage we might need to tune up the system for faster recommendation generation time.

REFERENCES

[1] G. Adomavicius, and A. Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." Knowledge and Data Engineering, IEEE Transactions on 17, no. 6 (2005): 734-749.

[2] S.B. Shirude and S.R Kohle, "A library recommender system using cosine similarity measure and ontology based measure", Advances in Computational Research,Vol. 4, Issue 1, 2012, pp. 91-94.

[3] A. Tejeda-Lorente , C. Porcel, E. Peis, R. Sanz, and E. Herrera-Viedma, "A quality based recommender system to disseminate information in a University Digital Library." Information Sciences (2013).

[4] P. Lakkaraju, S. Gauch, and M. Speretta, "Document similarity based on concept tree distance.", Proceedings of the nineteenth ACM conference on Hypertext and hypermedia. ACM, 2008, pp. 127-132

[5] M.J. Pazzani and D. Billsus, "Content-based recommendation systems", in: P. Brusilovsky, A. Kobsa, W. Nejdl (Eds.), The Adaptive Web, Lecture Notes in Computer Science, vol. 4321, Springer-Verlag, 2007, pp. 325–341.

[6] J.L. Herlocker, J.A, Konstan, L.G. Terveen, and J. Riedl, "Evaluating Collaborative Filtering Recommender Systems," ACM Transactions on Information Systems, 22(1), pp. 5-53, 2004.

[7] M. Balabanovic and Y. Shoham, "Combining content-based and collaborative recommendation", Comm. ACM, vol. 40, no.3, March 1997, pp. 66 -72.

[8] J. Beel, B. Gipp, S, Langer, and M. Genzmehr, "Docear: An Academic Literature Suite for Searching, Organizing and Creating Academic Literature", Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries (2011), 465–466.

[9] J. Beel, S. Langer, M. Genzmehr, and A. Nürnberger, "Introducing Docear's research paper recommender system", Proceedings of the 13th ACM/IEEE-CS joint conference, (JCDL '13), 2013, pp.459-460.

[10] J. Lee, K. Lee, and J. G. Kim, "Personalized Academic Research Paper Recommendation System.", arXiv preprint arXiv:1304.5457(2013).

[11] R. Torres, S. M. McNee, M. Abel, J.A. Konstan, and J. Riedl, "Enhancing Digital Libraries with Techlens", Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries (Tuscon, AZ, USA, 2004), pp. 228–236.

[12] B. Gipp, J. Beel, and C. Hentschel, "Scienstein: A Research Paper Recommender System", In Proceedings of the International Conference on Emerging Trends in Computing (ICETiC'09), Virudhunagar (India), January 2009, pp. 309–315.

[13] T. Strohman, W. B. Croft, and D. Jensen, "Recommending citations for academic papers.", In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '07, pages 705–706, New York, NY, USA, 2007. ACM.

[14] T. Bogers, and A. van den Bosch, "Recommending scientific articles using citeulike", Proceedings of the 2008 ACM conference on Recommender systems (2008), 287–290.

[15] J. Tang and J. Zhang, "A discriminative approach to topic-based citation recommendation", In T. Theeramunkong, B. Kijsirikul, N. Cercone, and T.-B. Ho, editors, Advances in Knowledge Discovery and Data Mining, volume 5476 of Lecture Notes in Computer Science, pages 572–579. Springer Berlin / Heidelberg, 2009.

[16] C. D. Manning, R. Prabhaka and H. Schütze, Introduction to information retrieval, New York: Cambridge University Press, 2008

[17] J. Sadiku and M. Biba, "Automatic Stemming of Albanian Through a Rule-based Approach", Journal of International Research Publications: Language, Individuals and Society, Vol. 6, 2012

[18] Nikitas N. Karanikolas, "Bootstrapping the Albanian information retrieval.", In: BCI'09. Fourth Balkan Conference in Informatics, pp. 231-235. IEEE, 2009

[19] A. Kadriu, "NLTK tagger for Albanian using iterative approach", Information Technology Interfaces (ITI), Proceedings of the ITI 2013 35th International Conference on , vol., no., pp.283,288, 24-27 June 2013

[20] B. Hasanaj, A Part of Speech Tagging Model for Albanian An innovative solution, Saarbrücken: LAP LAMBERT Academic Publishing, 2012.

[21] D. M. Blei, "Probabilistic topic models." Communications of the ACM 55, no. 4 (2012): 77-84

[22] A. Spahiu, "100 fjalët më të shpeshta në gjuhën shqipe", May 2010, <http://www.shkenca.org/pdf/gjuhe/100_fjale.pdf>

[23] C. Nascimento, A. H. Laender, A. S. da Silva, and M. A. Gonçalves. "A source independent framework for research paper recommendation." In Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, pp. 297-306. ACM, 2011.

[24] R.Kern, K. Jack, M Hristakeva, and M. Granitzer, "TeamBeam Meta-Data Extraction from Scientific Literature.", D-Lib Magazine 18, no. 7, 2012

[25] C. Giles, K. Bollacker, and S, Lawrence. "CiteSeer: An automatic citation indexing system.", Proceedings of the third ACM conference on Digital libraries. ACM, 1998.

On an Overlaid Hybrid Wire/Wireless Interconnection Architecture for Network-on-Chip

Ling Wang, Zhihai Guo, Peng Lv
Dept. of Computer Science and Technology
Harbin Institute of Technology
Harbin, China

Yingtao Jiang
Dept. of Electrical and Computer Engineering
University of Nevada, Las Vegas
Las Vegas, USA

***Abstract*—Network-on-Chip (NoC) built upon metal low-k interconnect wires, are to meet the ever stringent performance requirements in the future technology nodes. In response to this interconnection crisis, the wireless network-on-chip (WNoC), enabled by the availability of miniaturized on-chip antennas and transceivers, is envisioned one of the most revolutionary promising approach alternatives. In this paper, we present a new WNoC architecture with a layered topology, where a metal/low-k based wired network is partitioned into several subnetworks, and these subnetworks are connected through a wireless network that is overlaid on top of them. Due to limited transmission range, the wireless nodes in the wireless network actually communicate with each other in a multiple-hop fashion. As a large volume of traffic will go through the wireless nodes, a contention avoidance routing algorithm is adopted. In addition, two virtual channels have been introduced into the wireless router design to avoid any possible deadlocks that otherwise may occur. Experiment results have shown that throughput of the proposed architecture, on average, is about 20% higher than that of the existing WNoC architectures. And delay of the proposed architecture is about 30% less than the existing WNoC architectures.**

Keywords*—Network-on-Chip; wireless; subnet; 2-Level Hybrid Mesh topology

I. INTRODUCTION

Network-on-Chip(NoC) has emerged as a communication backbone to enable a high degree of integration in multi-core System-on-Chips (SoCs) [1]. Conventional NoCs rely on multi-hop packet-switched communications, where a data packet needs to pass through a series of routers/switches with considerable power and latency implications. To overcome these problems, express virtual channels are introduced to various NoC architectures [2~6] that can improve NoC's power, latency and throughput performance [2]. However, as the interconnection wires in these schemes [2-6] are still metal, delays of these metallic interconnects, governed by the physical law of showing a quadratic relationship with respect to the wire length, and are still quite long even for a modest routing distance. Therefore, on-chip interconnects carrying signals across different components will be the bottleneck to system performance and reliability, especially when CMPs scale to hundreds or thousands of cores on a chip. According to the International Technology Roadmap for Semiconductors (ITRS) [7], the wiring delay will be one of the critical issues of future designs.

The performance of NoC is expected to be significantly enhanced if wireless communication on chip technologies, such as Optical NoC, UWB and CMOS RF, are adopted [13-15]. The Optical and RF NoC are capable of inserting single-hop communication links between distant cores and thereby significantly reduce latency and power dissipation. On the other hand, the design of a wireless NoC based on CMOS ultra wideband (UWB) technology involves multi-hop communication through the on-chip short-range wireless channels. The performance of silicon integrated on-chip antennas for intra- and inter-chip communication with longer ranges have already been demonstrated in [11][12]. Antenna used in [13] can achieve a transmission range of only 1 mm but with a quite large size, the length is up to 2.98 mm. Consequently, for a NoC in a large die, say 20 mm x 20 mm, multi-hop transmissions are necessary for through-chip communications over the wireless channels. Moreover, the overheads of a wireless link may not be justifiable for 1 mm range of on-chip communication as compared to a wired channel.

In light of these technology advancements, the latest research is geared towards the mixed WNoC architectures which employ wired links between adjacent nodes and use one-hop or multi-hop wireless links between a few selected distant nodes [9][10][8]. Current WNoC architectures fall into two categories: single hop wireless NoC with long range on-chip wireless data links [9, 10], and multi-hop wireless NoC with short range on-chip wireless data links and larger number of wireless routers[8].

For one-hop wireless NoC architectures, data contention can cause severe performance problems at the wireless routers. The concept of subnet was first introduced in Small World WNoC [9], where nodes in a local subnet is wire linked and each subnet communicates with other subnets through a hub. This idea is inherited in Hybrid Mesh WNoC [10] where a traditional 2D Mesh is divided into several subnets, and each subnet has a wireless node in the center that allows this subnet to directly communicate with other subnets wirelessly. On the other hand, the multi-hop wireless NoC architectures [8], due to their higher number of wireless routers, can reduce the competition at each wireless router, but suffer from great power consumption and require large chip area.

To overcome the problems of these existing architectures, we propose a novel wireless WNoC architecture built upon two logically connected meshes, one wireless mesh and

another wired one. The wireless mesh supports multiple-hop wireless communications, so the routing paths in the wireless mesh are increased to avoid data congestion. We demonstrate the proposed architecture has low delay and high scalability.

The remainder of this paper is organized as follows. In section 2, we present the proposed WNoC architecture, followed by the description of the routing algorithm for the proposed architecture in Section 3. The wireless router design is detailed in Section 4. Section 5 presents the simulation results of the proposed architecture and its routing algorithm. Finally, we conclude the paper in Section 6.

II. WNOC ARCHITECTURE TOPOLOGY

To overcome many problems inherent in wired NoC, we propose a 2-Level WNoC architecture. Our WNoC architecture is based on a conventional wired 2-D mesh topology. Each IP here consists of a functional core, Network Interface (NI) and a router. Each router directly connects with its neighbor routers through multi-bit bidirectional links.

The proposed 2-Level WNoC architecture is shown in Fig. 1. In the lower wired mesh, the network is divided into a number of subnets. In each subnet, one wireless router (WR) is located in the center for inter-subnet wireless communication, and other wired routers are around the WR for intra-subnet wired communication. Then, all the WRs are connected to each outer by wireless links and constitutes the upper wireless mesh. Due to availability of multiple channels, Frequency Division Multiple Access (FDMA) method is adopted for channelization that can achieve simultaneous multiple communications between WRs.

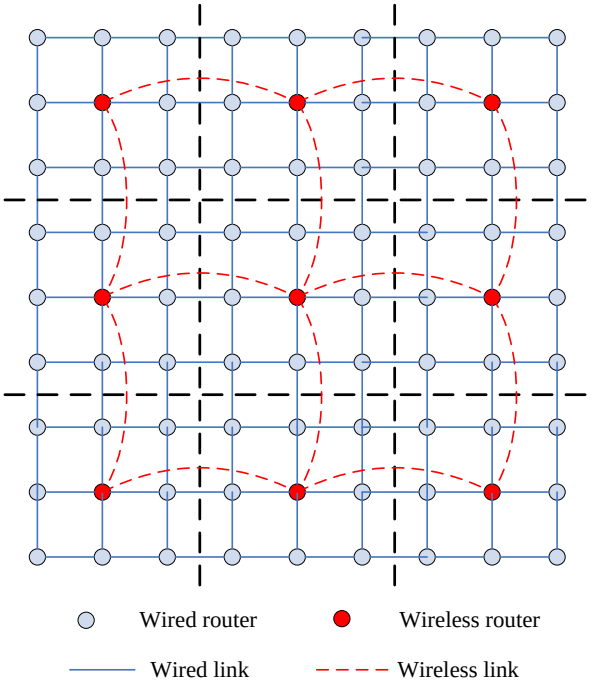


Fig. 1. Proposed WNoC topology

This architecture has the following properties:

In this structure, wireless nodes in the network are uniformly distributed and wireless data communications can pass multiple wireless hops (routers).

In this way, the wireless links will be less likely to be congested. Each subnet of the architecture has a fixed size, and the network is scalable.

Fig.2 depicts the two-level communications of a wireless router. Through the wired links in the lower mesh, the router can connect to its four neighbors of the E, S, N, W directions. The router is also connected to distant routers wirelessly in the X+, X-, Y+, Y- directions. When source PE starts communication, packets are injected into the network and routed through either wired mesh or wireless mesh.

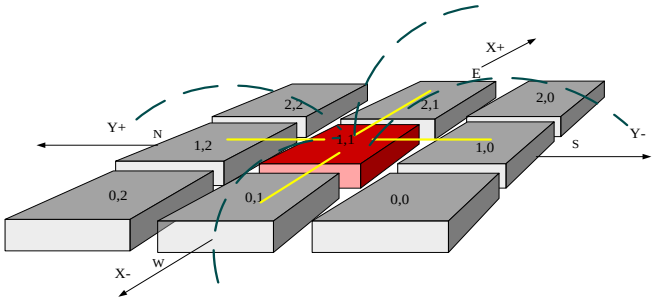


Fig. 2. Choosing the central node in a subnet

Comparisons of the proposed 2-Level WNoC with Hybrid Mesh and Small World are shown in Fig. 3 and Fig. 4. In terms of average distance of network, the 2-Level Hybrid Mesh topology is a little longer than Small-World, but shorter than Hybrid Mesh. However, when considering the number of wireless links, our proposed topology is more than the other two architectures gives more paths to route packets and balances the traffic of network. Hence, it contributes to reduce the probability of congestion in the upper wireless mesh.

III. ROUTING ALGORITHMS

The routing algorithm in the lower-layer wired mesh can be quite simple, while in the upper-layer wireless mesh, the routing algorithm needs to handle massive data volume passing through the wireless nodes.

For the proposed wireless NoC architecture, we design a routing algorithm (WFXY) with partial adaptiveness and congestion control:

Packets routed in the wired mesh follow the deterministic XY routing algorithm, which has a low algorithm complexity and guarantees the shortest path length;

In the top wireless mesh, the partially adaptive West-First routing algorithm is used to route packets to avoid data congestion.

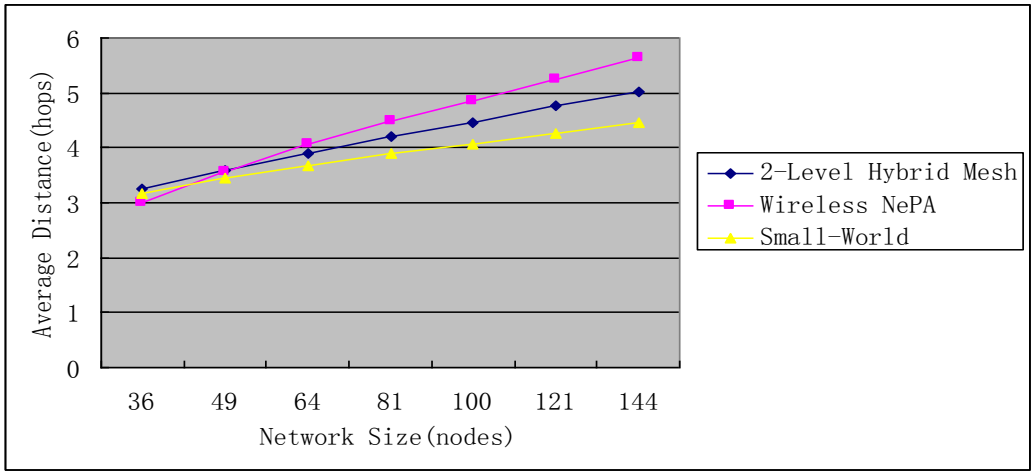


Fig. 3. Average Distance of the Three Topologies

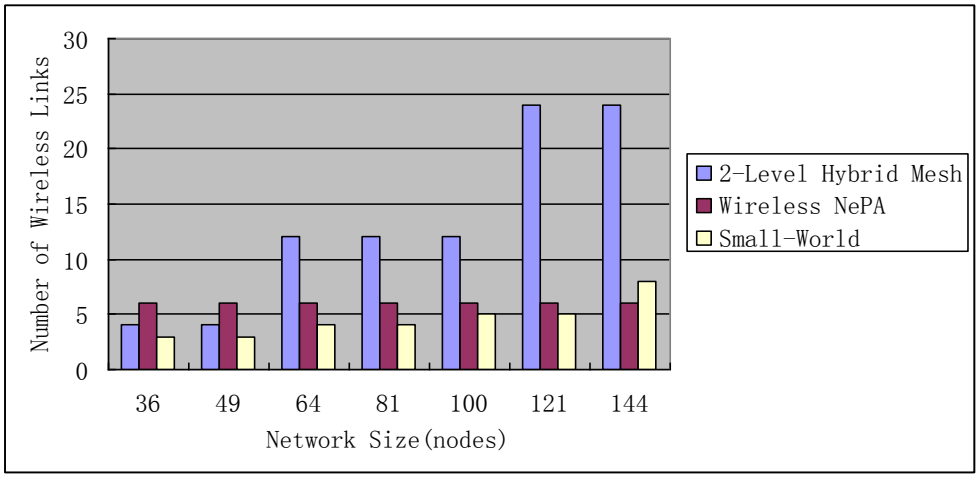


Fig. 4. Number of Wireless Links in the Three Topologies

Further, we define a threshold T for the routing distance. If a packet whose Manhattan distance between the source and the destination is greater than T, it will be classified as a long distance packet; otherwise, it is a short distance packet. The long distance packets are routed through wireless mesh, while the short distance packets can only be transferred through the wired mesh. In our experiment, the threshold T is set as 10.

A. *WFXY Routing Algorithm*

WFXY routing algorithm is a combination of West-First routing algorithm and XY routing. As a distributed routing algorithm, WFXY is implemented at every router, and the routing decision is made collectively by all the routers on the path from the source to the designation.

When a packet arrives at a node, WFXY algorithm will choose one from all the 8 directions (Figure 5) to switch the packet. This decision is based on the current node C, the destination node D, the packet type and available buffer sizes of its neighbor nodes.

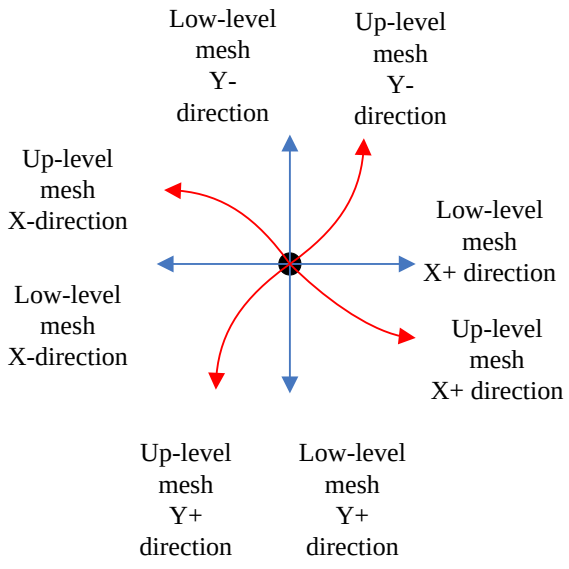


Fig. 5. 8 possible direction of the next hop

WFXY routing algorithm in wired nodes

Input: Current node C(Xc, Yc) and destination node D(Xd, Yd), the packet type;

Output: Packet routing decision;

Compute the subnet locations of nodes C and D;

If packet type ==0 then //--it is a short distance packet

Route the packet to D through the wired mesh using XY strategy;

Else //--it is a long distance packet

If C and D are in the same subnet then

Route the packet to D through wired mesh using XY strategy;

Else //--nodes C and D are in different subnets

Route the packet to the central node of current subnet through the wired mesh using XY strategy;

WFXY routing algorithm in wireless nodes

Input: Current node C(Xc, Yc) and destination node D(Xd, Yd), the packet type, the available buffer sizes of wireless nodes in neighbour subnets;

Output: Packet routing decision;

Compute the subnet locations of nodes C and D;

If packet type ==0 then //--it is a short distance packet

Route the packet to D through wired mesh using XY strategy;

Else //--it is a long distance packet

If C and D are in the same subnet then

Route the packet to D through wired mesh using XY strategy;

Else //--nodes C and D are in different subnets

Route the packet to the central node of destination subnet, through wireless mesh using West-First adaptive strategy;

B. *WFXY Algorithm analysis*

WFXY algorithm is a distributed algorithm that computing the next hop at every node takes time O(1). Hence, the overall time of determining a routing path is proportional to the length of the path. In our proposed architecture, the network diameter is smaller than 2N, so the total time complexity of WFXY is O (N). Moreover, because each node in the network keeps its location information for routing computation, the space complexity of WFXY is O (N^2).

IV. Wireless Router Design

In the previous section, we classify the data packets into two types: long distance packets and short distance packets. When these two kinds of packets exist in the network at the same time, it is very likely to cause a deadlock with formed cyclic routing paths involving both wired and wireless links. To resolve this potential deadlock problem, we introduce virtual channels into the router design.

The wireless router has 9 input ports, 5 wired ports and 4 wireless ports, while the wired router has 5 input ports, all wired ports, as shown in Figure 6. Each wired port can receive both types of packets with two virtual channels, VC0 and VC1. VC0 is for the long distance packets and VC1 for the short distance packets. As the wireless port handles long distance packets only, one buffer (VC0) is sufficient. The Switch Allocator handles the requests of the virtual channels, and the switch is used alternately by these VCs. Because the long distance packets and the short distance packets are routed through different virtual channels, and no VC can dictate the switching fabric indefinitely, the possibility of having a deadlock can be eliminated.

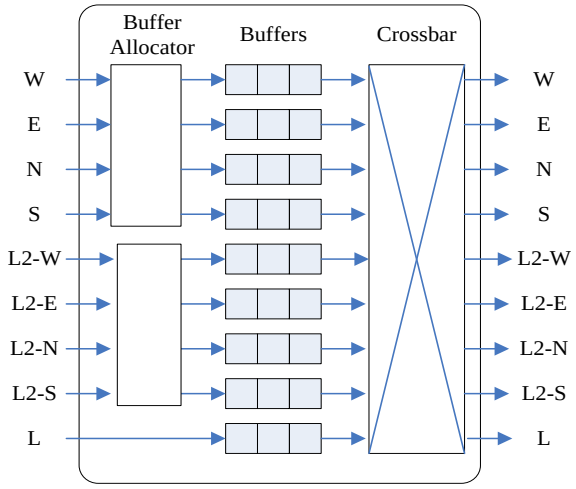


Fig. 6. The design of wireless router

Wired and Wireless Routers are designed by Verilog HDL. The synthesis for wire router and wireless router are realized by EDA tool ISE 10.1 of Xilinx Company. The type of FPGA is XC4VSX35. The synthesis result is shown in the table 1. As the number of ports of wireless router is more than that of wired router, the hardware overhead of wireless router is 1.6 times of that of wired router.

When the network has 144 cores, the area overhead for different WNoC architectures are shown in table 2. In Wireless NePA, the number of wireless routers is fewest, its area overhead is also lowest. As for Small-World, it needs some extra Hubs for inter-subnet communication, so it occupies largest area. Compared to these two architectures, the area overhead of our 2-Level Hybrid Mesh is quite modest, only 5% larger than that of Wireless NePA and 4.1% smaller than that of Small-World, respectively.

TABLE I. SYNTHESIS RESULT OF WIRED ROUTER AND WIRELESS ROUTER

Router	The number of Slices	The number of Slice Flip Flop	The number of LUTs
Wired Router	471	75	827
Wireless Router	752	131	1365

TABLE II. AREA OVERHEAD FOR DIFFERENT WNOC ARCHITECTURES

	2-Level Hybrid Meh	Wireless NePA	Small-World
Area Overhead (mm2)	10.912	10.396	11.36

V. PERFORMANCE ANALYSIS

To evaluate the performance of our proposed WNoC, a cycle-accurate WNoC simulator based on SystemC is used. In the experiment, we compare the performance (latency and throughput) of the proposed 2-Level Hybrid Mesh structure with that of two other WNoC architectures, Small-World and Wireless NePA. We assume that all three architectures are used to connect a system with 144 cores. Two traffic models are adopted in the experiment: (1) Uniform Random model, where every source node has equal probability to communicate with all other nodes; and (2) Hotspot model, where 8 hotspot nodes are introduced and they have to handle 15% of the total network traffic. The simulation environment is given in the table 3.

TABLE III. SIMULATION ENVIRONMENT SETTING

Parameter	Setting Result
Network Size	144 nodes
Buffer Size of Virtual Channel	16 Flits
Length of Flit	32 Bits
Length of packet	5 Flits
Simulation delay	20000 clock cycles
Timer cycle	4ns
Throughput model	Stochastic model、hotspot model

In the proposed WFXY routing algorithm, threshold T is set for dividing long distance and short distance packet according to the packet routing distance. When T is smaller, most of data packets in the network are classified as long distance packet. All of them routing at the up-level wireless Mesh network cause congestion at the up-level network. When T is larger, most of data packets in the network are classified as short distance packets. Short distance packets routing at the low-level lead to the increase of routing distance. Thus, throughput distribution at the two-level network is decided by the value of T. As threshold T has important impact on the performance of the network, the simulation experiment is used to decide the optimal value of T.

In the Figure 7, the network average delay versus T is given. Uniform stochastic model is adopted as throughput model. The injection rate is set as 0.3 and 0.4, respectively. When T is set as 8, more long distance packets routing at the up-level wireless mesh cause the congestion. Then the average delay is large. With the increase of T, the number of long distance packet is decreased. The congestion is alleviated, and then the delay is increased. Until T is set as 11, the least delay is achieved. With the increase of T, most of packets are classified as short distance packet routing at the low-level wired Mesh. Therefore the average routing distance of data packet becomes longer. And the average delay of network is increased gradually.

The average latency and throughput under Uniform Random pattern, measured against the traffic injection rate, are shown in Fig. 8 and Fig. 9, respectively. At low traffic load, all three architectures perform well. When the injection rate rises, the 2-Level Hybrid Mesh structure has the lowest latency and the highest throughput. It is shown in the Fig.9, when the injection rate is 0.3, average latency of the proposed architecture is lower than that of small world and wireless NePA by 50% and 27%, respectively. It is also shown in Fig.9, when the injection rate is 0.3, throughput of the proposed architecture is higher than that of small world and wireless NePA, 10% and 5%, respectively.

Under the Hotspot pattern, a lot of packets are transmitted to the 8 hotspot nodes, so the network is more likely to become congested. Fig.10 shows the average latency under the Hotspot model. When the injection rate is 0.3, average latency of the proposed architecture is less than that of small world and wireless NePA, 43% and 35%, respectively.

Thus, it is shown that the 2-Level Hybrid Mesh architecture has the lowest latency. In the 2-Level Hybrid Mesh WNoC, as multi-hop wireless links are used to transmit packets, the data can be scattered in different wireless paths. Moreover, the adaptive routing algorithm “West-First” is introduced in our design, which can degrade the congestion level of the network. So the 2-Level Hybrid Mesh WNoC has the lowest latency under heavy traffic load.

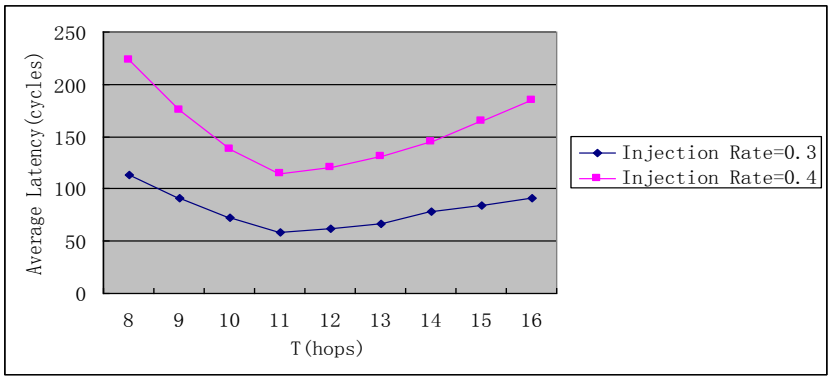


Fig. 7. Average Delay vs. Threshold value, T

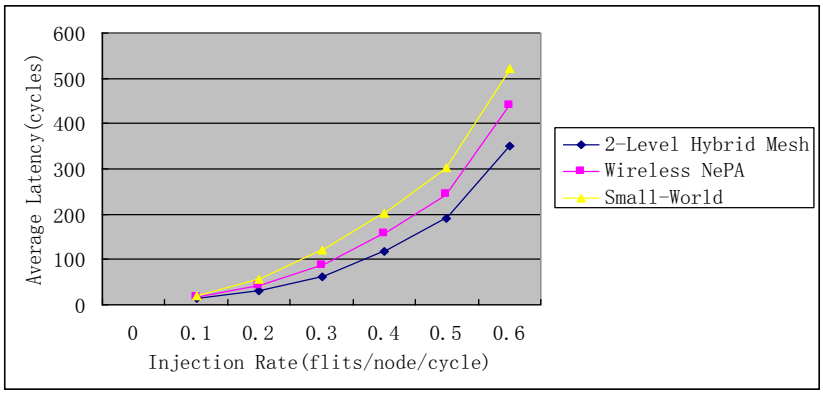


Fig. 8. Average Latency under Uniform Random model

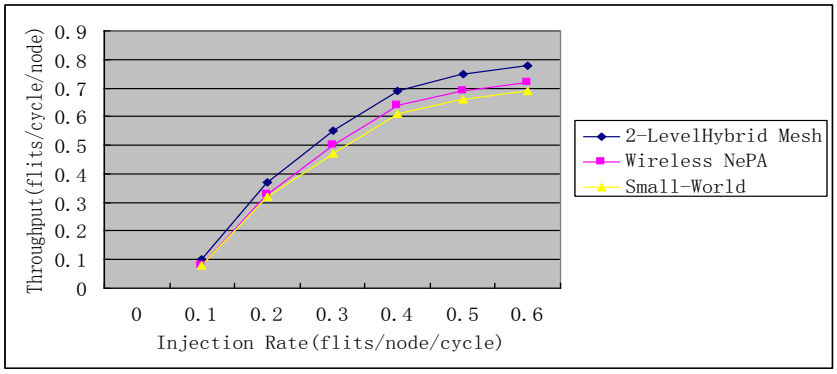


Fig. 9. Throughput under Uniform Random model

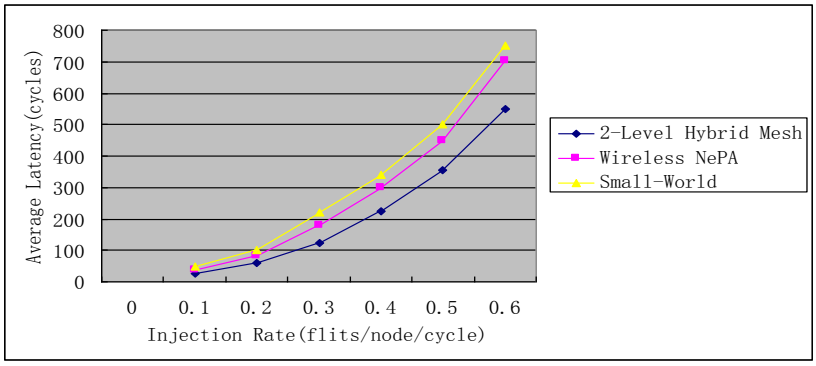


Fig. 10. Average Latency under Hotspot model

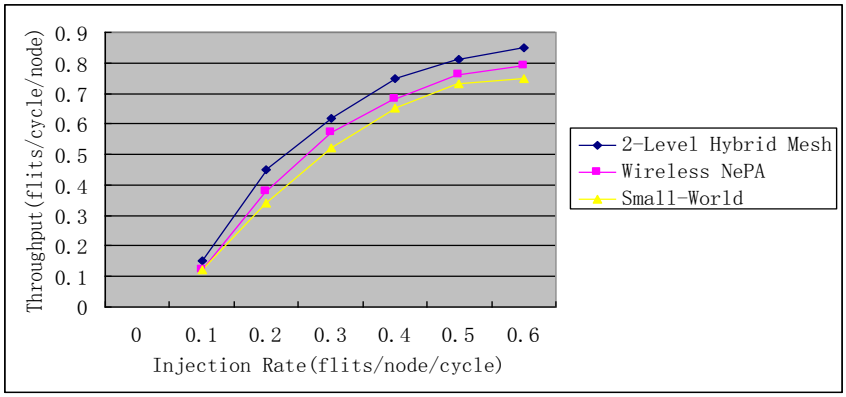


Fig. 11. Throughput under Hotspot model

In Fig. 11, the throughput under Hotspot pattern is shown. When the traffic is ligh with an injection rate of 0.3, throughput of the proposed architecture is higher than that of small world and wireless NePA, by 11% and 6%, respectively. Similar to the result under Uniform Random model, the 2-Level Hybrid Mesh architecture still has the highest throughput, proving it performs better than the other two architectures.

VI. Conclusion

In this paper, we have proposed a new WNoC structure, its routing algorithm, and correspondingly, the design of the wireless router. In essence, the proposed architecture is an overlay of two networks. At the upper layer, nodes can communicate through a wireless mesh network. While at the lower level, nodes can communicate by wired links. To avoid network congestion, packets are classified as long distance packets and short distance packets, and these two packets will be routed at different virtual channels in the upper wireless network to avoid any possible deadlocks. Experiment results have shown that the proposed NoC outperforms the other two existing WNoC architectures.

References

[1] A. Jantsch and H. Tenhunen (Eds.). Networks on Chip. Kluwer, 2003.

[2] P. P. Pande, C. Grecu, M. Jones, A. Ivanov and R. Saleh, "Performance evaluation and design trade-offs for network-on-chip interconnect architectures," IEEE Transactions on Computers, vol. 54, no.8, pp. 1025-1040, August 2005.

[3] Q. Yang and Z. Wu, "An improved mesh topology and its routing algorithm for NoC," International Conference on Computational Intelligence and Software Engineering (CiSE), pp.1-4, 2010.

[4] M. Saneei, A. Afzali-Kusha and Z. Navabi, "Low-latency multi-level mesh topology for NoCs," International Conference on Microelectronics (ICM'06), pp.36-39, 2006.

[5] K. Chen, C. Peng and F. Lai, "Star-type architecture with low transmission latency for a 2D mesh NOC," IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), pp.919-922, 2010.

[6] A. Tavakkol, R. Moraveji and H. SarbaziAzad, "Mesh connected crossbars: Anovel NoC topology with scalable communication bandwidth," International Symposium Parallel and Distributed Processing with Applications (ISPA), pp 319-326, 2008.

[7] International Technology Roadmap for Semiconductors, 2007 edition

[8] D. Zhao, Y. Wang, J. Li and T. Kikkawa, "Design of multi-channel wireless NoC to improve on-chip communication capacity," IEEE/ACM International Symposium on Networks on Chip (NoCS), pp. 177-184, 2011.

[9] S. Deb, K. Chang, A. Ganguly and P. Pande, "Comparative performance evaluation of wireless and optical NoC architectures," IEEE International on SOC Conference (SOCC), pp.487-492, 2010.

[10] C. Wang, W. Hu and N. Bagherzadeh, "A wireless network-on-chip design for multicore platforms," 19th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP), pp.409-416, 2011.

[11] J. Lin, H. Wu, Y. Su and L. Gao, "Communication using antennas fabricated in silicon integrated circuits, " IEEE Journal of Solid-state Circuits, vol. 42, no. 8, pp. 1678-1687, August 2007.

[12] S. B. Lee, S. W. Tam, I. Pefkianakis and S. Lu, "A scalable micro wireless interconnect structure for CMPs," ACM Annual International Conference on Mobile Computing and Networking (MobiCom), pp. 20-25, September, 2009.

[13] A. Shacham and K. Bagman, "Photonic network-on-chip for future generations of chip multi-processors," IEEE Transactions on Computers, vol. 57, no. 9, pp. 1246-1260, Sept. 2008.

[14] M. F. Chang, J. Cong, A. Kaplan and M. Naik, "CMP network-on-chip overlaid with multi-band RF-interconnect," IEEE International Symposium on High-Performance Computer Architecture (HPCA), pp. 191-202, 16-20 February, 2008.

[15] D. Zhao and Y. Wang, "SD-MAC: design and synthesis of a hardware-efficient collision-free QoS-aware MAC protocol for wireless network-on-chip,"IEEE Transactions on Computers, vol. 57, no. 9, pp. 1230-1245, September 2008.

Image Sharpness Metric Based on Algebraic Multi-Grid Method

Qian Ying , Ren Xue-mei, Huang Ying, Meng Li
Laboratory of Graphics Image and Multimedia,
Chongqing University of Posts and Telecommunications
Chongqing, China

***Abstract*—In order to improve Mean Square Error of its reliance on reference images when evaluating image sharpness, the no-reference metric based on algebraic multi-grid is proposed. The proposed metric first reconstructs the original image by Algebraic Multi-grid (AMG), then compute the Mean Square Error between original image and reconstructed image, the result represents image sharpness. Experiments show that the proposed sharpness metric has better practicability and monotonicity, correlates well with the perceived sharpness. The algorithm has superiority in image sharpness metric.**

***Keywords*—image sharpness mean square error; algebraic multigrid method; sharpness metric; image reconstruction**

I. INTRODUCTION

There has been an increasing need to develop quality measurement techniques that can predict perceived image/video quality automatically. These methods are useful in various image/video processing applications [1-5], such as compression, communication, printing, display, analysis, registration, restoration and enhance [6]. Subjective quality metrics are considered to give the most reliable results since, for many applications, it is the end user who is judging the quality of the output. Subjective quality metrics are costly, time-consuming and impractical for real-time implementation and system integration. On the other hand, objective metrics can be divided into three categories: full-reference, reduced-reference, and no-reference, which is the most convenient. The traditional sharpness metrics are gradient function, such as Sum Modulus Difference (SMD), Variance are gray scale function, and entropy function. In recent years, Marziliano[7] and Ong et al. measure the image based on smoothing effects of edge blur. Ferzli[8] put forward perceptual sharpness metric based on measured just-noticeable blurs (JNBs), but unable to keep balance between stability and sensitivity. Narvekar and Karam[9] estimate the sharpness of an image as the cumulative probability of detecting blur at an edge (CPBD). Mean square error (MSE) is a full-reference evaluation methods commonly used, which requires a reference to calculate sharpness of distortion image. In this paper, we propose an improved MSE together with reconstruction image use algebraic multi-grid. The proposed metric scans for the whole image. The clearer the image is, the smaller the similarity between pixels and the smaller of MSE between image reconstructed by algebraic multi-grid and Original Image. So the metric proposed could used to measure image sharpness.

II. PROPOSED NO-REFERENCE OBJECTIVE SHARPNESS METRIC

A. Algebraic multi-grid is an iterative method used for solving the matrix equation automatic and established on geometric multi-grid [10]. Algebraic multi-grid is mainly used for solving large-scale scientific project computation, especially partial differential equations (group). AMG is allowed to solve the non-structure mesh, and therefore it is more easily extended to image processing [11-12]. AMG is mainly applied in image reconstruction, binary, recovery and denoising [13-14]. When applied Algebraic multi-grid in image reconstruction, first, we should convert the image into graph, then, create relationship affinity (affinity) matrix on similarity between pixels gray value of image. The similarity between pixels can be calculated by weight function, the commonly function used is:

$$Wij = \exp(-\frac{\| F_i - F_j \|_2^2}{\sigma_l^2}) * \begin{cases} \exp(-\frac{\| x_i - x_j \|_2^2}{\sigma_x^2}) & if \ \| x_i - x_j \|_2 < r \\ 0 & otherwise \end{cases} \quad (1)$$

Where, F_i and F_j are pixels' gray values of point i and j in image, x_i is pixel's value of coordinates in space, σ_l is standard deviations of Gaussian function, σ_x is standard deviations of space coordinates function, r is striking distance between two nodes. From the weight function we know that the closer two points' distance and gray values are to each other, the greater the similarity of two points.

Secondly, extract coarsening sequence of image and define the original coefficient matrix as finest mesh, the original definition of the coefficient matrix of the finest mesh Ω_0.

In order to derive a coarse level system, we first need a splitting of Ω_m into two disjoint subsets $\Omega_m = F^m + C^m$, with C^m representing those variables which are to be contained in the coarse level (C-variables) and F^m being the complementary set (F-variables). According to the above description, we regard the set of coarse-level variables as a subset of fine-level ones. The coarse grid is a subset of its finer grid.

Generally, C^m and F^m are selected as follows

1) For any $i \in F^m$ and $j \in S_i^m$ *(strongly connected to* i *), we know that* $j \in C^m$ *or* j *is strongly connected to point in* C^m.

2) C^m *is the largest point set formed by strong connection point.*

Where the strong connection point is defined as follows:

$$-aij \geq \theta_0 \max_{k \neq i}(-a_{ik}), 0 < \theta_0 \leq 1 \quad (2)$$

This definition is actually for M-matrix, which is symmetric positive definite matrix and non-diagonal elements are no positive. θ_0 is usually taken 0.25.

The M-matrix is:

(1) $a_{i,i} > 0 \ \forall i$;

(2) $a_{i,j} > 0 \ \forall i \neq j$;

(3) $A^{-1} \geq 0$

Finally, we could get reconstructed images through interpolating image coarsening sequence. The interpolating algorithms most commonly used are nearest neighbor interpolation, bilinear interpolation and bicubic interpolation. The nearest-neighbor interpolation algorithm selects the value of the nearest point and does not consider the values of neighboring points at all, yielding a piecewise-constant interpolation. Bilinear interpolation an extension of linear interpolation for interpolating functions of two variables on a regular 2D grid. Bicubic interpolation is an extension of cubic interpolation for interpolating data points on a two dimensional regular grid. The algorithm not only considers the influence of 4 directly adjacent pixel gray value the surrounding pixel gray scale value of four, but the variance rate of gray level.

B. No-reference image sharpness metric based on AMG

MSE is a traditional full-reference objective image quality evaluation. The method is easy to calculate, but it just a pure mathematical statistic of pixels error without consideration for correlation between pixels. The no-reference image sharpness metric based on AMG is an improvement to MSE. The improved algorithm is no-reference, without non-distorted image and more real-time. Firstly, we process the target image to achieve the first layer coarsening sequence by AMG. Then the reconstructed image could get by interpolation method. The MSE of the original image the reconstructed image is used to measure the sharpness of image.

MSE is defined as follows:

$$M = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(f_{ij}^{'} - f_{ij})^2 \quad (3)$$

Where, i and j are pixel coordinates, f_{ij} is the original image, $f_{ij}^{'}$ is the distortion image of f_{ij}.

III. Results and Analysis

A practical evaluation must meet the following criteria:

1) Prediction Monotonicity: Image sharpness metric scores should show a corresponding increase and decrease monotonically as the image sharpness increases and decreases.

2) Prediction Consistency: A metric must perform well regardless of the Content of image it is given. A good indicator should always perform well with different image content.

3) Prediction Accuracy: This refers to the ability to correctly evaluate image quality, can generally be determined by the index of the MOS value for comparison.

A. Prediction Monotonicity

Test set: There are 6 512 × 512 house images, include one original picture and five blurred images using a lowpass7 × 7Gaussian mask with standard deviation σ equal to 0.4, 0.8, 1.2, 1.6 and 2.0, respectively, as shown in Fig. 2。From the results shown in Figure 2, we know that the more blur image becomes, the smaller the sharpness metric value. The algorithm meets monotonicity principles.

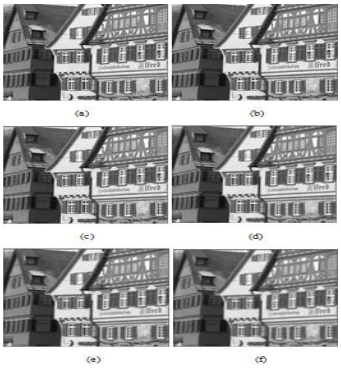


Fig. 1. 6 different ambiguity houses

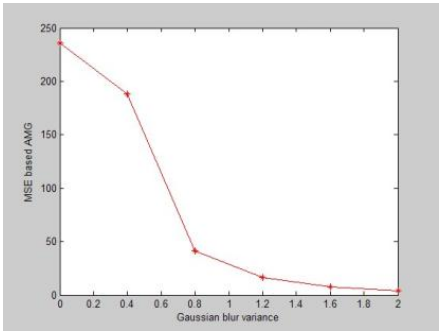


Fig. 2. Different sharpness value of 6 houses

B. Prediction Consistency

- First, we cut the lena image into 4 same area of four dimensions and adopt gauss filter to process the 4 images. For the 4 sub-pictures, we do as follows:

The 4 sub-pictures are blurred using a lowpass7 × 7Gaussian mask with standard deviation σ equal to 1, 2.5, 4, and 5.5, respectively, as shown in Fig. 3。The result is shown in Table below. From it we know that the algorithm performs well regardless of the Content of image it is given. The result is consistent with its blurring.

TABLE I. STANDARD DEVIATION 1、2.5、4、5.5

image name	std_dev	Proposed Metric
(a)	1	34.99856
(b)	2.5	9.674969
(c)	4	5.505095
(d)	5.5	2.263391

- Secondly, we metric the 4 images sharpness by the proposed algorithm, Tab.2 shows the result. And Tab.3 shows the metric result of images, which we give a Gaussian filtering conflict to result of Tab.2 with a lowpass7 × 7Gaussian mask of standard deviation σ equal to 5.5, 2.5, 4 and 1.

TABLE II. STANDARD DEVIATION 5.5、2.5、4、1

image name	std_dev	Proposed Metric
(a)	5.5	1.61791
(b)	2.5	9.67496
(c)	4	5.50509
(d)	1	36.3086

C. Prediction Accuracy

- To test the performance of the metric, all of Gaussian-blurred images from the LIVE[15]. Each image was rated by about 20–29 subjects. The subjects were asked to rate the images on a continuous linear scale which was divided into five different regions namely, “Bad,” “Poor,” “Fair,” “Good,” and “Excellent.” The raw scores for each subject were converted to difference scores and then z-scores. The scores were then scaled and shifted to a range of 1 to 100. Then the difference mean opinion score (DMOS) and mean opinion score (MOS) for each image was calculated. We use 6 kinds of algorithms to process on the 84 images from LIVE. To measure how well the proposed metric, the authors followed the suggestions of the VQEG report where several evaluation metrics are proposed. The predicted MOS values are then used in calculating the performance measures including PCC (Pearson correlation coefficient, indicates the prediction accuracy), SROCC (Spearman rank-order correlation coefficient, indicates the prediction monotonicity), RMSE (root mean squared prediction error), MAE (mean absolute prediction error) and OR (outlier ratio, indicates consistency) and Spearman correlation coefficients should be high and the values of RMSE, MAE, and OR should be low. The result is given below.

TABLE III. PERFORMANCE COMPARISON OF 6 ALGORITHMS

Metrics	Pearson	Spearman	RMSE	MAE	OR
CPBD	**0.908**	**0.930**	**0.095**	**7.518**	**5.766**
JNBM	0.829	0.808	0.191	10.052	7.691
SMD	0.721	0.792	0.298	12.456	9.391
entropy	0.218	0.214	0.512	17.539	14.867
Variance	0.106	0.239	0.5	17.869	15.204
Proposed Metric	**0.918**	**0.949**	**0.107**	**7.129**	**5.705**

It can be seen from tab.3, in the above 5 indicators, the proposed metric algorithm is better than JNBM and SMD. But for the CPBD, the proposed metric is bad in the RMSE index, mainly because of the large variation range of metric values in proposed metric. Monotonicity and accuracy of proposed Metric is higher than that of CPBD.

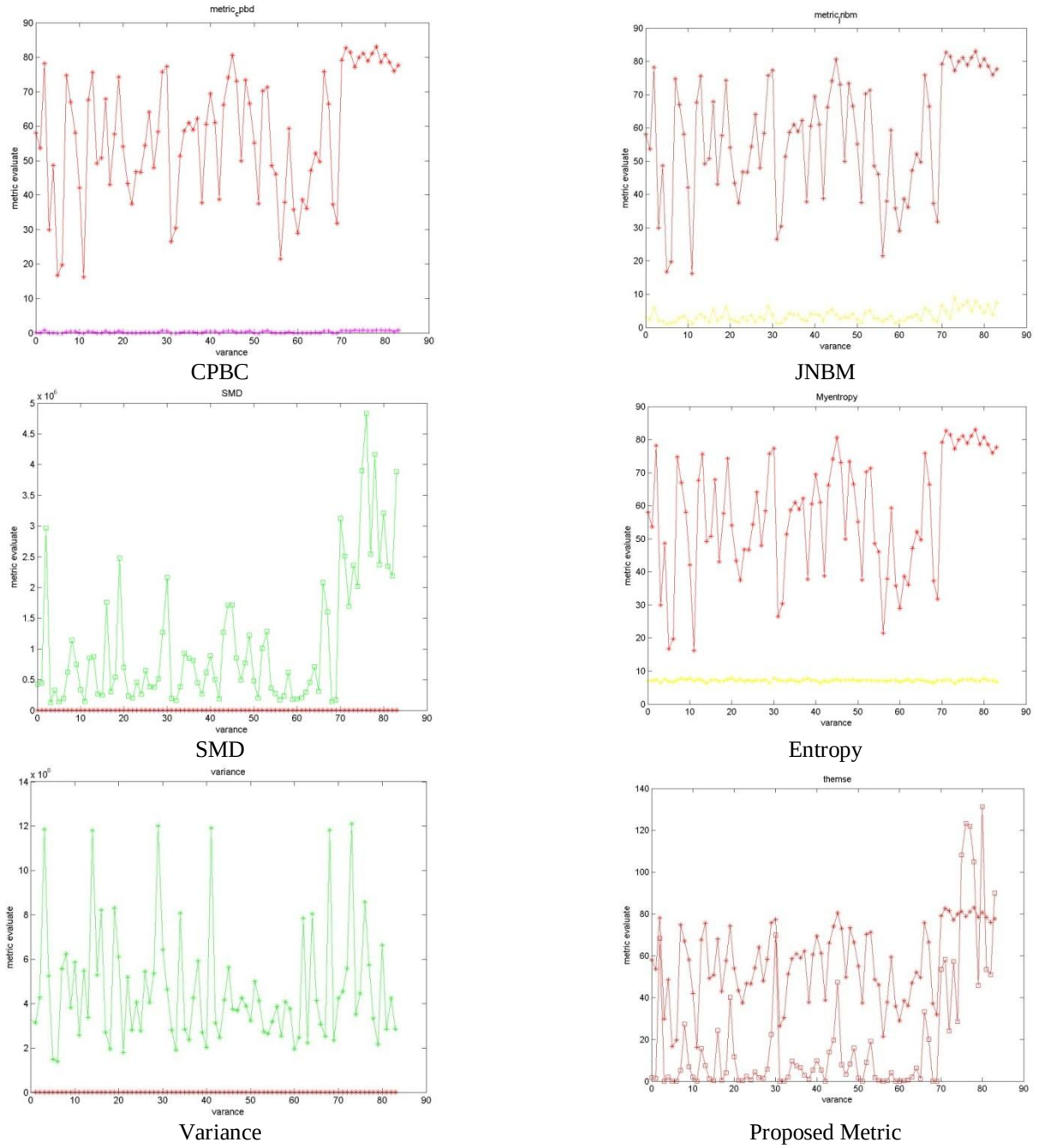


Fig. 3. Comparation of 6 algorithms and the MOS values

Fig.3 gives the fitting curve of 6 algorithms and the MOS values. From fig.2, we know that the proposed metric is good fitting to the MOS values. Where, the red line with symbol '*'indicates MOS values.

IV. CONCLUSION

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper.

In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper.

REFERENCES

[1] Wee C Y, ''Paramesran R. ''*Image sharpness measure using eigenvalues*[C]'', 9th International Conference on Signal Processing, 2008: 840-843.

[2] Zhu and P. Milanfar, "*A no-reference sharpness metric sensitive to blur and noise*," in 1st International Workshop on Quality of Multimedia Experience (QoMEX), 2009.

[3] Crete F, Dolmiere T, Ladret P, et al. ''*The blur effect: perception and estimation with a new no-reference perceptual blur metric*[C]''. International Society for Optics and Photonics, 2007: 64920I-64920I-11.

[4] Shaked D, Tastl I. ''*Sharpness measure: Towards automatic image enhancement*[C]'', IEEE International Conference on Image Processing. IEEE, 2005, 1: I-937-40.

[5] Hassen R, Wang Z, Salama M. ''*No-reference image sharpness assessment based on local phase coherence measurement''* C, 2010

IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010: 2434-2437.

[6] Li Tan, Zhengguo Li, Yih Han Tan, Rahardja, S."*Relevant MSE-Based Image Quality Metric,"J*. IEEE Transactions on Image Processing. vol.22 , no. 11, pp. 447-4459, 2012

[7] Feichtenhofer, C. ; Fassold, H. ; Schallauer, P. "*A Perceptual Image Sharpness Metric Based on Local Edge Gradient Analysis*," J.Signal Processing Letters, IEEE, VOL.20, NO.4, PP.379 - 382,2013.

[8] R. Ferzli, L. J. Karam. "*A no-reference objective image sharpenss metric based on the notion of just noticeable blur (jnb)*," J, IEEE Transactions on Image Processing, vol.18, pp.717–728, 2009

[9] N. D. Narvekar, L. J. Karam. " ," J, IEEE Transactions on Image Processing, vol.20, no.9, pp. 2678–2683, 2011

[10] R. D. Falgout. "*An Introduction to Algebraic Multigrid," J, Computing in Science & Engineering*,"vol.8, no.6, pp.24-33, 2006.

[11] Kimmel R, Yavneh I. "*An algebraic multigrid approach for image analysis*, " J,SIAM, vol. 24, no.4, pp. 1218-1231, 2003.

[12] Vaněk P, Mandel J, Brezina M. ''*Algebraic multigrid by smoothed aggregation for second and fourth order elliptic problems*''J. Computing, 1996, 56(3): 179-196.

[13] Chan T F, Saad Y. ''*Multigrid algorithms on the hypercube multiprocessor''*J. Computers, IEEE Transactions on, 1986, 100(11): 969-

[14] Xu Qiubin. "*Numericals for total variation-based reconstruction of motion blurred images*," J, Applied Mathematics-a Journal of Chinese University, vol.25, no.3, pp. 367-373,2010.

[15] Yiping Xu, Hanlin Chen, Kelong Zheng. "*A Combination Algorithm for Image Denoising and Deblurring*," J, Wireless Communications Networking and Mobile Computing(WiCOM), 2010, pp.1-4.

Investigating Students' Achievements in Computing Science Using Human Metric

Ezekiel U. Okike
Department of Computer Science
University of Botswana
Gaborone, Botswana

***Abstract*—This study investigates the role of personality traits, motivation for career choice and study habits in students' academic achievements in the computing sciences. A quantitative research method was employed. Data was collected from 60 computing science students using the Myer Briggs Type indicator (MBTI) with additional questionnaires. A model of the form $y_{ij}=\beta_0+\beta_1 x_{1j+}\beta_2 x_{2j+}\beta_3 x_{3j+}\beta_4 x_{4j+}\ldots\beta_n x_{nj}$ was used, where y_{ij} represents a dependent variable, $\beta_0+\beta_1 x_{1j+}\beta_2 x_{2j+}\beta_3 x_{3j+}\beta_4 x_{4j+}\ldots\beta_n x_{nj}$ the independent variables. Data analysis was performed on the data using the Statistical Package for the social sciences (SPSS). Linear regression was done in order to fit the model and justify its significance or none significance at the 0.05 level of significance. Result of regression model was also used to determine the impact of the independent variable on students' performance. Results from this study suggest that the strongest motivator for a choice of career in the computing sciences is the desire to become a computing professional. Students' achievements especially in the computing sciences do not depend only on students temperamental ability or personality traits, motivations for choice of course of study and reading habit, but also on the use of Internet based sources more than going to the university library to read book materials available in all areas**

Keywords—academic achievement; personality traits; computing science; study habits

I. INTRODUCTION

Achievements in educational terms refer to academic achievement. It is the performance of a student in his studies at school. Student's achievement in school subjects such as Mathematics, Physics, and Computer Science is a measure of the overall academic ability and knowledge of a subject of study. Although there exists a number of achievement studies in the subject areas like Mathematics, Physics, Chemistry and Biology, this is not the case in Computer Science especially at the university level. The need to measure students achievement in computing science among other things include the following: to ensure that students meet their set academic goals, to ensure students meet graduation requirements, to serve as a means to validate teaching effectiveness, and to serve as a means to identify outstanding students for recognition.

Every university takes as priority the learning standards and outcomes of her students. Hence universities adopt different approaches to measuring students academic achievement. A common approach used by many universities to measure academic achievement of students is by means of Continuous Assessment (CA), and final examination. In this regard, the CA mark could be between 30%-40% while the final examination score could be 60%-70%. Furthermore, universities have tools that help in ascertaining how well a subject has been taught by a lecturer and how well the students understood and mastered the course content. An example of this tool at the University of Botswana is the Students Evaluation of Courses and Teaching (SECAT) tool. Using this tool, the course, the student and the course lecturer are evaluated by the students through an automated questionnaire which reports its analysis as soon as students complete the SECAT questionnaires. Although the use of this tool is a good way to measure how well a course has been taught by a lecturer, and how well the students have mastered the course content, there remains a gap to be investigated between the students inherent personality trait and students achievement in each course of study. This paper investigates students achievement in Computing Science using 60 third year students of Computer Science (CS), Information Technology (IT), Computing with Finance, and Information System (IS) at the University of Botswana, Gaborone. This study is motivated by the interest to contribute to the empirical body of knowledge about using a human metric tool such as the Myers Briggs Type Indicator (MBTI) as a predictor of students' achievement especially in the Computing Sciences.

The term Computing Science encompasses Computer Science (CS), Computer Engineering (CE), Software Engineering (SE), Information Technology (IT), and Information Systems (IS). For the purpose of this study, courses offered by students in the Department of Computer Science leading to the award of Bachelor of Computer Science (BSC 280), Bachelor of Information Technology (BSC 204), and Bachelor of Computing with Finance (BSC 205) and Bachelor of Information Systems (BIS 230) are considered. All courses offered in these programmers cover hardware and software courses representing the four subdivisions of Computing Science as defined by the educational curriculum committee of the professional body in charge of computing education worldwide [1].

A. Problem Statement

The use of a human metric tool such as the Myers Brigg Type Indicator (MBTI) to predict academic achievement in the computing sciences has not been widely reported. In effect,

there is not enough empirical evidence as to the role of personality traits in students' academic achievements especially at the tertiary level. This study is a contribution to bridge the gap in literature regarding academic achievements in computing sciences using the MBTI tool.

B. Study Ovjectives

The main objective of this study is to investigate if personality traits do affect academic achievements in computing science. The study also investigates the motivating factors affecting the choice of a career in Computing Science and the reading habits which influence academic achievements in computing science

C. Research Questions

The following research questions are investigated in this study.

- What are the factors that influence students choice of course of study in Computing Science at the University of Botswana?
- Which study habits influence students academic success in the Computing Science at the University of Botswana?
- Which personality traits are high achievers in Computing Science at the University of Botswana?

D. Research Hypotheses

The following hypotheses are tested in this study:

H0: Introverts will have higher academic achievements than extroverts

H1: Introverts will not have higher academic achievements than extroverts

H0: Sensors will have higher academic achievements than intuitives

H1: Sensors will not have higher academic achievements than intuitives

H0: Thinkers will have higher academic achievements than feelers

H1: Thinkers will not have higher academic achievements than feelers

H0: Judges will have higher academic achievements than Perceivers

H1: Judges will not have higheracademic achievements than perceivers

H0: There is significant correlation between personality traits and academic achievements

H1: There is no correlation between personality traits and academic achievements.

The rest of this paper is divided into 6 sections. Section 2 is a review of relevant literature. Section 3 explains the research methodology. Section 4 presents the result of this study with appropriate discussion. Section 5 is the conclusion while section 6 is the list of references

II. Literature Review

Okike[2] investigated the major personality indicators of students Systems Analysts and Designers at the University of Botswana and their performances at System Analysis and Design practical and theoretical examinations. The study suggests that the best achievers in Systems Analysis and Design are students who possess the personality types of Extrovertism (E), iNtuition (N), Feeling (F), Judging (J), Thinking (T), Introversion (I), and Sensing (S). The highest passes in the overall Systems Analysis and Design examination are students with the combined personality traits of Introversion iNtuition Feeling Judging (INFJ), Introversion iNtuition Thinking Judging (INTJ), Extroversion iNtuition Thinking Judging (ENTJ), ExtroversioniNtuition Feeling Judging (ENFJ), and Introversion Sensing Thinking Judging (ISTJ).

Capretz and Ahmed [3] studied the connections between personality traits and the process of software development. The authors mapped soft skills and personality traits to the main stages of the software life cycle. They claim that assigning people with personality types best suited to particular stages of the software life cycle increases the chances of project's successful outcome

Omar and Syed-Abdullah [4] applied rough setsin identifying effective personality type in software engineering teams. It was suggested that a balance of personality types Sensing (S), iNtuition (N), Thinking and Feeling (F) assisted teams in achieving higher software quality. Extroverts (E) in the team also had impacts on team performance.

Da Cunha and Greathead[5] investigated if a specific personality type is correlated with performance on code review task. In their investigation, the researchers measured personality with the Myers Briggs Type indicator(MBTI) while the reviewed code was a Java based 282 lines of code. The subjects of study were 64 second year undergraduate student at New Castle University, UK. To examine the possible links with MBTI type and code review ability, the researchers computed some correlations between task score and each bipolar factor Extrovert-Introvert (EI), Sensing-iNtuition(SN), Thinking-Feeling(TF) and Judging-Perceiving(JP). The result of this study indicated that only a single bipolar within the SN bipolar significantly correlated with code review task, suggesting that people more intuitively inclined performed better than others on code review.

Bishop-Clark and Wheeler [6] investigated the Myers-Briggs personality type and its relationship to computer programming. Specifically, the study sought to know if college students with certain personality types performed better than others in an introductory programming course. The researchers first did a pilot study with 24 students and a follow up study with 114 students. The result of this study showed that sensing students performed significantly better than intuition students in programming assignments while judging students performed better than perception students on computer programs although the results were not significant statistically. In addition, they also noted that although personality may not be an important factor in a student's decision to drop a course, it may influence a student's

evaluation of a class. The researchers concluded that the act of programming (creating and debugging programs) is a feat on its own and should be distinct from scores on written programs.

Similarly, Irani, Telg, Scherler, and Harrington [7] studied the relationship between personality type and distance education students course perception and performance using 39 graduate students of distance education. Perceptions of instructional technique used by the distance instructor were strongly correlated to the students' course grade and overall grade point average for the following personality types: extravert, introvert, intuitive, sensing, feeling, and judging. Of the MBTI type preferences, only thinking and perceiving types showed no significant correlations between course perceptions and performance indicators. Findings from this study indicate that performance outcomes for distance education students may be closely related to course perceptions as a function of personality type preference. Perceptions of instructional technique used by the distance instructor were strongly correlated to the students' course grade and overall grade point average for the following personality types: extravert, introvert, intuitive, sensing, feeling, and judging.

Turley and Bieman[8] studied the attributes of individual software developers in order to identify their professional competencies using biography data and Myers- Briggs Type Indicator (MBTI) and concluded that there was no simple predictor of performance. Although experience variables in their study were related to performance, it could only predict classification of exceptional and non exceptional of 63% of the subjects.

Chung [9] studied the cognitive abilities in computer programming using 523 Form Four secondary school students in Hong Kong. Test administered to the students included mathematics, space, symbols, hidden figures and programming ability. Results of the study suggest that performance in mathematics and spatial tests were significant predictors in programming ability.

III. Study Methodology

This study employs quantitative research methods. A human metric tool (Myers Brigg's Type Indicator, MBTI) and a supportive questionnaire were administered on 60 third year students taking the Bachelor of Information Technology (BSC204), Bachelor of Computing with Finance (BSC 205), Bachelor of Information Systems (BIS 230),Bachelor of Computer Science (BSC280) and Bachelor of Education, Computer Science option (BED 240) programmes of study at the University of Botswana.

The MBTI tool is an automated questionnaire based personality test or human metric tool which reports individual personality trait based on the 16 recognizable traits namely Introversion Sensing Thinking Judging (ISTJ), Introversion Sensing Feeling Judging (ISFJ), Introversion Sensing Thinking Perceiving (ISTP), Introversion Sensing Feeling Perceiving; Introversion iNtuition Feeling Judging (INFJ), Introversion iNtuition Thinking Judging (INTJ), Introversion iNtuition Feeling Perceiving (INFP), Introversion iNtuition Thinking Perceiving (INTP); Extraversion Sensing Thinking Perceiving (ESTP), Extraversion Sensing Feeling Perceiving (ESFP), Extraversion Sensing Thinking Judging (ESTJ), Extraversion Sensing Feeling Judging (ESFJ); Extraversion iNtuition Feeling Perceiving (ENFP), Extraversion iNtuition Thinking Perceiving (ENTP), Extraversion iNtuition Feeling Judgingg (ENFJ), and Extraversion iNtuition Thinking Judging (ENTJ). The MBTI tool classified the 60 students according to their individual personality traits. Furthermore, additional questionnaires were designed in order to gather information from the students concerning what motivated their choice of programme of study at the University of Botswana (UB): BSC 204, BSC 205, BSC 280, BIS 230 and BED 240; as well as how they study in order to understand the various courses they take at the university. In order to measure achievement in the various courses taken by a student, a model was used as follows:

$$y_{ij=}\beta_0+\beta_1 x_{1j+}\beta_2 x_{2j+}\beta_3 x_{3j+}\beta_4 x_{4j+}\ldots\beta_n x_{nj}$$

Where y_{ij} represents a dependent variable, and the independent variables are represented as $\beta_0 + \beta_1 x_{1j+} \beta_2 x_{2j+} \beta_3 x_{3j+} \beta_4 x_{4j+} \ldots \beta_n x_{nj}$. The dependent variable in this study is the achievement (performance or score) in each course taken by a student; the independent variables are the various personality traits exhibited by a student in terms of the level (in percentages) of Extroversion, Introversion, Thinking, Feeling, Sensing, iNtuition, Judging and Perceiving.

The variables which influenced students choice of programme of study: parental influence, personal desire to be in the computing profession, students ability in science and mathematics, students ability in science without mathematics, other reasons; and the variables which indicate student study habit : reading of text books, reading only class notes, reading from online lecture notes (module), use of internet materials, use of university library to read text books and other relevant materials, none use of university library, going to university library to read personal materials; reading class notes, text books and online lecture notes; reading class notes and online lecture materials only because student don't have enough money to purchase recommended text; any other reasons were also considered as independent variable. Data analysis was performed on the data using the Statistical Package for the social sciences (SPSS). Linear regression was done in order to fit the model and justify its significance or none significance at the 0.05 level of significance. Result of regression model was also used to determine the impact of the independent variable on students performance.

IV. Result and Discussion

A. Model Fitting

The model $y_{ij=}\beta_0+\beta_1 x_{1j+}\beta_2 x_{2j+}\beta_3 x_{3j+}\beta_4 x_{4j+}\ldots\beta_n x_{nj}$ was tested for fitness using regression statistic. The result is shown in Table 1 and Table 2. In Table 1, the R square value of 0.431implies that about 43.1% of the predictors explain the variations in the dependent variable. This means that the personality traits contribute to academic achievements of computing science students.

TABLE I. MODEL SUMMARY

Model	R	R Sqr	Adjustd R Sqr	Std Error of the estimate
1	.657[a]	.431	-.032	7.22360

TABLE II. ANOVA

Model 1	Sum of Sqr.	df	Mean Sq	F	Sig
Regrsion	1069.104	22	48.596	.931	.563[b]
Residual	1408.872	27	52.180		
Total	3430.733	59			

a) Dependent Variable: AVGSCORE

b) Predictors:(Constant), PERSONALITY TRAITS, MOTIVATION CHOICE OF STUDY, STUDY HABITS

In Table 2 (ANOVA), the model is not significant. However, considering Table 1, it is clear that 43.1% of the independent variables (predictors) explains the variation in the dependent variable (average score or achievement). This suggests that more information on some other variables are needed to predict achievement in the model

B. Analysis of Students by Personality types

Table 3 below shows the frequency distribution of students by personality types.

TABLE III. DISTRIBUTION OF STUDENTS BY PERSONALITY TYPE

		Frequency	Percent	Valid %
Valid	ENTJ	10	16.7	16.7
	ENFJ	10	16.7	16.7
	ENFP	1	1.7	1.7
	ESFJ	3	5.0	5.0
	ESFP	2	3.3	3.3
	ISFJ	4	6.7	6.7
	ISTJ	4	6.7	6.7
	ISTP	1	1.7	1.7
	INFJ	12	20.0	20.0
	INTJ	12	20.0	20.0
	ISFP	1	1.7	1.7
	Total	60	100.0	100.0

From Table 3, the highest personality traits among this set of students are Introversion iNtuition Feeling Judging (INFJ), Introversion iNtuition Thinking Judging (INTJ), Extroversion iNtuition Thinking Judging (ENTJ) and Extroversion iNtuition Feeling Judging (INFJ)

C. Factors Influencing Students choice of Coputing career

Table 4 shows the frequency distribution of students' response to their motivations for choice of course of studyin computing science.

TABLE IV. MOTIVATIONFORCHOICE OFCAREERINCOMPUTING SCIENCE

Variable	Yes	No	Mean
Parental influence	2 (3.3%)	57 (95.0 %)	60
Desire to be in computing profession	46 (76.6%)	13 (21.7%)	60
Ability in Science and Mathematics	20 (33.3%)	38 (63.3%)	60
Ability in Science , but fair in Mathematics	2 (3.3%)	57 (95.0%)	60
Other reasons	4 (6.7%)	55 (91.5%)	60

From Table 4, the desire to be in the computing profession is the highest motivating factor (76.6%) in choosing a career in computing science followed by students' ability in science and Mathematics (33.3%). Factors such as parental influence (3.3%), being good in science but not in Mathematics (3.3%) do not have a strong influence on students choice of course of study in Computing science. This suggests that ability in Mathematics is a significant predictor of a students probability of choosing a career in the Computing Sciences. Chung [9] also suggested that ability in Mathematics and spartial tests were significant predictors of programming ability and hence programming as a career. Hence, achievements in Mathematics could enhance achievements in Computing Science.

D. Students Study Habits in Computing Science

Table 5 presents students' study habit and understanding their course lectures in computing science at the University of Botswana.

From Table 5, reading class notes, text books and materials posted by lecturers on Moodle (76.6%) and use of Internet related sources (58.3%) are the main study habits of computing science students at the University of Botswana. In terms of students' use of the University library, 33.3 % of computing science students use the library in order to study their personal materials, 23.3% use library book materials while 6.7% of computing science students do not use the library.

TABLE V. STUDENTS STUDY HABITS IN COMPUTING SCIENCE

Variable	Yes	No	Mean
I read my text books	38 (63.3%)	20 (33.3 %)	60
I read my class notes only	14 (23.3%)	43 (71.7%)	60
I read my notes from Moodle only	16 (26.7%)	42 (70.0%)	60
I use only Internet materials	35 (58.3%)	23 (38.3%)	60
I use the library to read text books	14 (23.3%)	43 (71.7%)	60
I don't use the library	4 (6.7)	54(90.0%)	60
I use the library to read personal materials	20 (33.3%)	38 (63.3%)	60
I read class notes, text books & Moodle materials	46 (76.6%)	12 (20.7%)	60
I read class notes and Moodle materials only	0	58 (96.7%)	60
Other reasons	1(1.7 %)	57 (95.0%)	60

Furthermore, it is interesting to note that students' use of Internet based sources (58.3%) is higher than students use of the University library to study personal materials (33.3%), and to use library book materials (23.3%). This suggests that Computing Science students spend more time using Internet based sources than using the University library. This implies that computing science students use the laboratories more than the library.

TABLE VI. ACHIEVEMENT SUMMARY OF NUMBER OF PASSES BY PERSONALITY TRAITS

COURSES \ PERSONALITY & SCORE	**ENTJ**	**ENFJ**	**ENFP**	**ESFJ**	**ESFP**	**ISFJ**	**ISTJ**	**ISTP**	**ISFP**	**INFJ**	**INTJ**
No in class	10	10	1	3	2	4	4	1	1	12	12
Discrete Structure 1 Passes											
50-69%	4	6	1	2	1	2	3	0	1	9	9
70-100%	5	4	0	1	0	0	1	1	0	3	3
Discrete Structure 2											
50-69%	7	7	0	1	1	2	3	0	1	7	7
70-100%	2	3	0	0	0	0	0	1	0	1	3
Algebra											
50-69%	6	6	0	2	1	1	3	1	157		
70-100%	2	3	0	1	0	0	1	0	157		
Programming Principles											
50-69%	9	8	0	3	2	4	4	0	0	9	9
70-100%	1	2	1	0	0	0	0	1	1	3	3
O.O Programming											
50-69%	4	3	1	2	0	1	3	0	2	7	0
70-100%	5	7	0	0	0	1	0	1	1 83		

Data Structures														
50-69%	7	9	1	1	0	3	4	0	07	11				
70-100%	1	1	0	0	0	0	0	1	120					
Data Base Systems														
50-69%	7	3	1	3	2	3	4	1	0811					
70-100%	3	7	0	0	0	1	0	0	14	0				
General Computing (intro)														
50-69%	7	8	1	3	1	2	4	1	1	10	10			
70-100%	2	2	0	0	0	0	0	0	0	1	0			
Total Passes														
No of A's	21	29	1	2	0	2	2	3	52719					
No of Passes	51	50	5	17	8	18	28	3	66264					

E. Achievement of Computing Science students by Personality Type

From Table 6, achievement in 6 core Computing Science courses by personality type is presented. The core courses are taken by students offering Computer Science, Computing with Finance, Information Technology and Information Systems. For the course Discrete Mathematics I, the best students are those who possess personality types ENTJ with 5As, ENFJ with 4As, INFJ with 3As and INTJ with 3As. For Discrete Mathematics II, the best students are of the personality types INTJ with 3As, ENFJ with 3As, ENTJ with 2As, ISTP with 1A, and INFJ with 1A. In Algebra, the best students are students with the personality types INTJ with 7 As, INFJ with 5As, ENFJ with 3As, ENTJ with 2As, ESFJ with 1A, ISTJ with 1A and ISFP with 1 A.

In Programming Principles, the best students are of the personality types INFJ with 3As, INTJ with 3As, ENFJ with 2As, ENTJ with 1A, ISTP with 1A, and ISFP with 1A. In Object Oriented Programming, the best students possess the personality types INFJ with 8As, ENFJ with 7As, ENTJ with 5As, INTJ with 3As, ISTP with 1A, and ISFP with 1A. In Data Structures, the best students possess the personality types INFJ with 2As, ENTJ with 1A, ENFJ with 1A, ISTP with 1A, and ISFP with 1A. In Data Base Systems, the best students possess the personality types ENFJ with 7As, INFJ with 4As, ENTJ with 3As, ISFJ with 1A, and INTJ with 1A. In General Computing (Introduction to Computing), the best students possess the personality types ENTJ with 2As, ENFJ with 2As and INFJ with 1A. For all courses, 'A' grades range from 70% to 100%. Overall results of achievement from Table 6 indicate that the highest number of achievers in a prioritized order are students who possess the personality traits Extroversion iNtuition Feeling Judging (ENFJ) with 29 'A' grades, Introversion iNtuition Feeling Judging (INFJ) with 27'A' grades, Extroversion iNtuition Thinking Judging (ENTJ) with 21 'A' grades, Introversion iNtuitionThinking Judging (INTJ) with 19'A' grades, Introversion Sensing Feeling Judging (ISFP) with 5'A' grades, Introversion Sensing Thinking Perceiving (ISTP) with 3 'A' grades, Introversion Sensing Thinking Perceiving (ISTP) with 3 'A' grades, Introversion Sensing Thinking Judging (ISTJ) with 2 'A' grades, Introversion Sensing Feeling Judging (ISFJ) with 2 'A' grades, Extroversion Sensing Feeling Judging (ESFJ) with 2 'A' grades, and Extroversion iNtuition Feeling Perceiving (ENFP) with 1 'A' grade.

F. Comparisons Between Personality characteristics and Computing Characteristics

a) Characteristics of various types

From [2], the following characteristics of the various personality types were identified:

- Extraversion (E): Focus on the outer world
- Introversion (I): Focus own inner world
- Feeling (F): When making decisions, they look at the people and special circumstances
- iNtuition (N) : Interpret and add meaning to information they taken in
- Judging (J): In dealing with outside, they get things decided
- Thinking (T): When making decisions they first look at the logic and consistency

b) Essential skills of Computing Scientists

Grimson [10] identified four skill set required to study Computer Science namely:

- Computational thinking skill
- Understanding Code
- Understand abilities and limits
- Map Problem into computation.

Capretz and Ahmed[3] on the other hand identified thesoft skills requirements of Systems Analysts, Software Designers, Programmers, Testers and Maintenance engineers and subsequently mapped the skills unto personality types. Some of the identified soft skills include communication skills, interpersonal skills, ability to work independently, being an active listener, having strong analytical and problem solving skills, being open and adaptable to changes, innovative skills, organizational skills, acute attention to details, fast listening skills and team playing skills:

G. Discussion

A careful comparison of Tables 3 and 6 suggests that the dominant personality type among the students are Introversion iNtuition Feeling Judging (INFJ), Introversion iNtuition Thinking Judging (INTJ), Extroversion iNtuition Feeling Judging (ENFJ) and Extroversion iNtuitionThinking Judging (ENTJ). Of the dominant personality types, ENFJ presents the best achievers with 29 'A' grades, followed by the types INFJwith 27 'A' grades, ENTJ with 21 'A' grades, INTJ with 19 'A' grades and ISFP with 5 'A' grades. Overall, the highest passes in the 8 core courses considered in this study are INFJ (89 passes) and INTJ (81 passes) supporting hypotheses 1,3,4, 5 (bullets); ENFJ (79 passes) and ENTJ (72 passes) which nullifies hypothesis 2 (bullet); ESFP(3passes), ISFJ (3passes), ISTJ (3 passes), ESFP (2 passes), ENFP (1), ISTP (1), ISFP (1).

H. Conclusion

In conclusion, the desire to be in the Computing profession is the essential motivating factor in choosing a career in the Computing Sciences (research question 1).

The main study habits which influence students achievement in Computing Science are reading class notes, text books, materials posted on Moodle by lecturers and use of Internet related sources (research question 2). The use of Internet related sources imply that Computing Science students spend more time in laboratories with computers on Internet facilities than in than library reading book sources. Distinctive personality types may enhance academic achievement as well as performance in certain tasks. It is suggested that the best achievers from this study are students who possess the combine personality types ENFJ, INFJ, ENTJ, INTJ (research question 3). Therefore, study concludes that personality traits do affect achievement in Computing Science and especially the traits of Extroversion, Introversion, iNtuition, Feeling, Thinking, Judging, and probably Sensing and Perceiving.

REFERENCES

[1] ACM/AIS/IEEE-CS."Computing Curricular 2005".

[2] E. U. Okike. "Bipolar Factor and Systems Analysis Skills of Student Computing Professionals at University of Botswana, Gaborone," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 5. No. 3, 2014

[3] F.Capretz, F. Ahmed, "Making sense of software development and personality types," ITPro, vol. 12, no. 1 January/February 2010.

[4] M. Omar, and S Syed-Abdullah, "Identifying effective software engineering (SE) team personality types composition using rough set approach," IEEE 2010

[5] D. A. Da Cunha, and D. Greathead, " Does Personality matter? An Analysis of code –review ability," Communications of the ACM. Vol. 50. No. 5, pp. 109-111, May 2007

[6] C. Bishop-Clark and D. Wheeler, "The Myers-Briggs personality type and its relationship to computer programming," Journal of Research on Computing in Education. Vol. 26 Issue 3, pg. 358-371, 1994.

[7] T. Irani, R. Telg, C. Scherler, and M Harrington, " Personality type and its relationship to distance education students' course perceptions and performance," Quarterly Review of Distance Education; Vol. 4 Issue 4, p445, 2003.

[8] R. T. Turley, and J. M. Bieman, "Competencies of exceptional and non exceptional software engineers," J. Systems Software. 28:19-38

[9] C. Chung , "Correlates of problem solving in programming," CUHK Educational Journal Vol. 16. No. 2, pp.185-190, 1986

[10] E. Grimson."Introductory Computer Science Lecture Notes on MIT OpenCourseware,"unpublished

Malware Detection in Cloud Computing

Safaa Salam Hatem
College of Science,
University of Al-Qadisiyah,
Al-Qadisiyah, Iraq

Dr. Maged H. wafy
Information Technology department,
Faculty of Computers and
Information, Helwan University,
Cairo, Egypt

Dr. Mahmoud M. El-Khouly
Information Technology department,
Faculty of Computers and
Information, Helwan University,
Cairo, Egypt

***Abstract*—Antivirus software is one of the most widely used tools for detecting and stopping malicious and unwanted files. However, the long term effect of traditional host based antivirus is questionable. Antivirus software fails to detect many modern threats and its increasing complexity has resulted in vulnerabilities that are being exploited by malware. This paper advocates a new model for malware detection on end hosts based on providing antivirus as an in-cloud network service. This model enables identification of malicious and unwanted software by multiple detection engines Respectively,This approach provides several important benefits including better detection of malicious software, enhanced forensics capabilities and improved deployability. Malware detection in cloud computing includes a lightweight, cross-Storge host agent and a network service. In this paper Combines detection techniques, static signatures analyze and Dynamic analysis detection. Using this mechanism we find that cloud- malware detection provides 35% better detection coverage against recent threats compared to a single antivirus engine and a 98% detection rate across the cloud environment.**

Keywords— Malware; Security; Cloud Computing

I. INTRODUCTION

Detecting malicious software is a complex problem. The vast, ever-increasing ecosystem of malicious software and tools presents a daunting challenge for network operators and IT administrators. Antivirus software is one of the most widely used tools for detecting and stopping malicious and unwanted software. However, the elevating sophistication of modern malicious software means that it is increased challenging for any single vendor to develop signatures for every new threat. Indeed, a recent Microsoft survey found more than 45,000 new variants of backdoors, Trojans, and bots during the second half of 2006 [1].

In this paper, we suggest a new model for the detection functionality currently performed by host-based antivirus software. This paper is characterized by two key changes.

- *Malware detection as a network service*: First, the detection capabilities currently provided by host-based antivirus software can be more efficiently and effectively provided as an in-cloud network service. Instead of running complex analysis software on every end host, we suggest that each end host runs a lightweight process to detect new files, send them to a network service for analysis, and then permit access or quarantine them based on a report returned by the network service.
- *Multi-detection techniques*: Second, the identification of malicious and unwanted software should be determined by multiple, Different detection engines Respectively. Suggest that malware detection systems should leverage the detection capabilities of multiple, Collection detection engines to more effectively determine malicious and unwanted files.

In the future, we will see an increase in the dependence of cloud computing as consumers increasingly move to mobile platforms for their computing needs. Cloud technologies have become possible by tuberculation in order to share physical server resources between multiple virtual machines (VMs). The advantages of this approach include an increase in the number of clients that can be served for every physical server and the ability to provide software as a service (SaaS).

In this paper, previous work on malware detection had been presented, both conventional and in the presence of cloud as storage in order to determine the best approach for detection in the cloud [2]. We also argue the benefits of multiple detection throughout the cloud and present a new approach to coordinate detection across the cloud.

Section II provides background and related work the research area, specifically: cloud technologies, security system in the cloud, malware detection and detection in the cloud. Section III, we explain our Proposed System. Section IV we show Remarks of our system. Finally, section V Conclusions the points raised in this paper and provide some ideas for future work.

II. BACKGROUND

A. Cloud Computing

With the Internet's ubiquity in modern living, many argue that some level of cloud computing is now a common occurrence. This research heavily focuses on cloud computing technology, and thus requires a formal definition of cloud computing. Cloud computing cannot be easily defined. There are many definitions, which share the same common denominator: the Internet. Cloud computing is a way to use the Internet in the daily life of a single machine or single room, using all the tools installed on computers [Figure 1]. It is also the ability to use shared computing resources with local servers handling applications. With cloud computing users do not worry about the location and the storage of their data. They just start using the services anywhere and at any time. The main driver of this technology is Virtualization (Hypervisor) and virtual appliance [3]

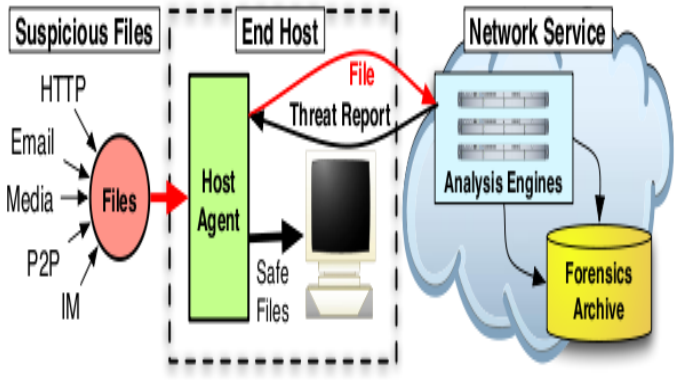


Fig. 1. The Flow of the Process of the cloud computing systems

Cloud computing offers different service models that allow customers to choose the appropriate service model that fits their environment needs, Cloud service models are software as a service (SaaS), Platform as a service (PaaS), and Infrastructure as a service (IaaS) [4] [5]:

- Software-as-a-service (SaaS): The consumer uses the provider's applications, which are hosted in the cloud. For example, Salesforce.com CRM Application.
- Platform-as-a-service (PaaS): Consumers deploy their own applications into the cloud infrastructure. Programming languages and application development tools used must be supported by the provider. For example, Google Apps.
- Infrastructure-as-a-service (IaaS): Consumers are able to provide storage, network, processing, and other resources, and deploy and operate arbitrary software, ranging from applications to operating systems.

B. Related Work

As a matter of fact "cloud computing" concepts date backward to the 1950s, when large-scale mainframes were made available to schools and corporations, In addition, the on-demand computing concept of the cloud model went back to the time-sharing era in the 1960s [7]. Therefore, many of the cloud computing security issues are arguably quite similar to the ones that were introduced during the Internet expansion era. However, Malware detection in a Cloud Computing service was explicitly introduced in [8], what we now commonly refer to as cloud computing is the result of an evolution of the widespread adoption of Virtualization, service-oriented architecture, autonomic, and utility computing. Details such as the location of infrastructure or component

Devices are unknowns to most end-users, who no longer need to be thorough, understand or control the technology infrastructure that supports their computing activities. There are several previous studies related to this research dealing with all of cloud computing and its structure as well as detection systems used for each of the Static analysis, detection: Signature Optimizing Pattern Matching and Dynamic analysis detection: Heuristic, Can be summarized as follows:

C. In Cloud Computing

Oberheide [9] proposed in his thesis "N-Version Antivirus in the Network Cloud" a new model for antivirus deployment by providing antivirus functionality Asa network service using N-version protection. This novel paradigm provides significant advantages over traditional host-based an antivirus, including better detection of malicious software, enhanced forensic's capabilities, improved deployable and manageable retrospective detection. Use a production implementation and real-world deployment of the Cloud AV platform. In addition, Schmidt, et. Al [10], presented an approach for combined malware detection and kernel rootkit prevention in virtualized cloud computing environments, and all running binaries in virtual instance are intercepted and submitted to one or more analysis engines. Besides a complete check against a signature database, lives introspection of all system calls is performed to detect yet unknown exploits or malware.

Malware detection has been an important issue in computing since the late '80s. Since then the predominant method of malware detection has been to scan a computer system for infection by matching malware signatures to files on the computer. Although detection of known samples is extremely reliable, signature based detection only works for malware that has been obtained, analyzed and a suitable signature identified. Murad et al. [11], showed that signature based detection can be thwarted by analyzing the malware instructions and identifying the instructions that comprise the signature.

Li [6] decreased the signature mapping cost by optimizing signature library, taking advantage of common conduct characteristics of viruses such as self-replicate and seasoning, and proposed optimization policy against this scalability issue with the help of data mining. Moreover, he decreased the number of unnecessary signature matching and raises efficiency of that comparison procedure by rearrangement within a signature library. In Heuristic detection, Treadwell [12] suggested analyzing the obfuscation pattern before unpacking, providing a chance to prevent malware from further execution. In this paper, we propose a heuristic detection approach that targets obfuscated windows binary files being loaded into memory.

III. Proposed System and Results

This paper proposes a malware detection system to be built on cloud environment,

Initially, we will divide the system architecture into two main sections according to the mechanism of action of each part. First Section, relating cloud computing and the second section, explains the two detection techniques that used. A cloud computing, we use cloud as software as a service (SAAS) which is a new service and an information delivery model that utilizes existing technologies [14]. The proposal of this work is to find the optimal solutions to the problems of anti-viruses and improve performance and find possible alternatives for a better working environment without problems with high efficiency and flexibility. In this system, a traditional detection technique as per static signatures and dynamic detection technology has been used. Then, safer

system methods and modern to rival existing anti-virus has been selected, for this a hybrid system of two detection methods has been created:

1) Static analysis: Signature Optimizing Pattern Matching.

2) Dynamic analysis: Heuristic.

Both of them will be explained below.

a) Signature Optimizing Pattern Matching: This method is used depends on the signature, which storage already in the database. For this purpose, we used a string matching algorithm, comparison variants of which arise in finding similar DNA or protein sequences.

It is important to use this method to our system Because of there are several new viruses detected; therefore, it becomes necessary to add their signatures to a library. To this end, a failure comparisons increase, this would negatively affect the efficiency of the signature matching procedure. Based on the virus characteristic of self-replicating and seasoning, this system proposed optimizing policy focus on Signature library; one common feature of virus is that it will scan targeted files and inject the malicious code into the normal files. So lots of replicas coexist within one system. So when any virus is detected by signature match, this virus signature is temporarily stored, so the other replicas do not need to match against the other large amount of signatures in the actual signature library [Figure 2].

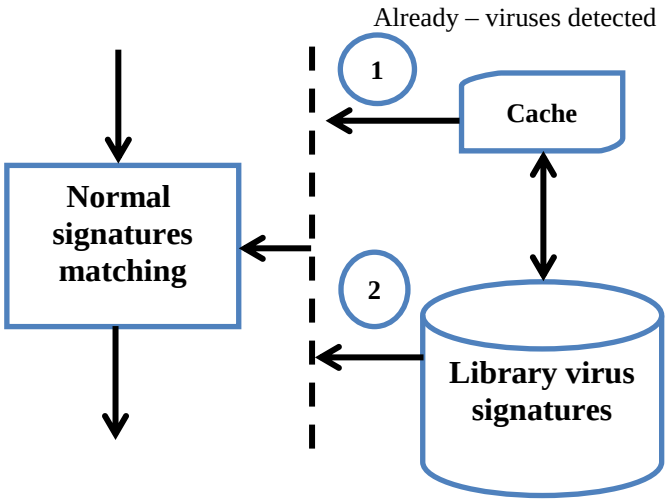


Fig. 2. Shows the process for Optimizing Pattern Matching of Library signatures

So this pre-comparison with already-detected viruses will reduce the signature matching times. [6]

b) Heuristic Detection: antivirus software often used one or several techniques proactively detect malware. This method is dependent on analyzing suspicious file's characteristics and behavior to determine whether it is indeed a malware, Heuristic analyzer (or simply, a heuristic), i.e. A technology in virus detection, which cannot be detected by antivirus databases. It allows detecting objects, which are suspected as being infected by unknown or new modification of known viruses. Files which are found by the heuristic analyzer are considered to be probably infected. [7]

In addition, an analyzer usually begins by scanning the code for a suspicious attributes (commands) characteristic of malicious programs. This method is called static analysis. For example, many malicious programs search for executable programs, open the files found and modify them.

A heuristic examines an application's code and increases its "auspiciousness counter" for that application if it encounters a suspicious command. If the value of the counter after examining the entire code of the application exceeds a predefined threshold, the object is considered to be probably infected.

Moreover ,We connect these processes and initial database in a cloud computing environment to be lighter weight , speed processing and performance; we used a file transfer protocol (FTP) For this purpose to connect between the database system in the cloud and the internal processes of the detection system .

We used the technique of detected in real time (RTP) to detect any suspicious attack on real time for working, In addition, sending notifications to the user in the end -host if there an attack or suspecting files,

Thus the user is using the action required for Eliminate or fix. When finding cases of suspected, unknown virus Signature automatically added to our database system. Figure 3, shows the simple outline of Mechanisms in the proposed system.

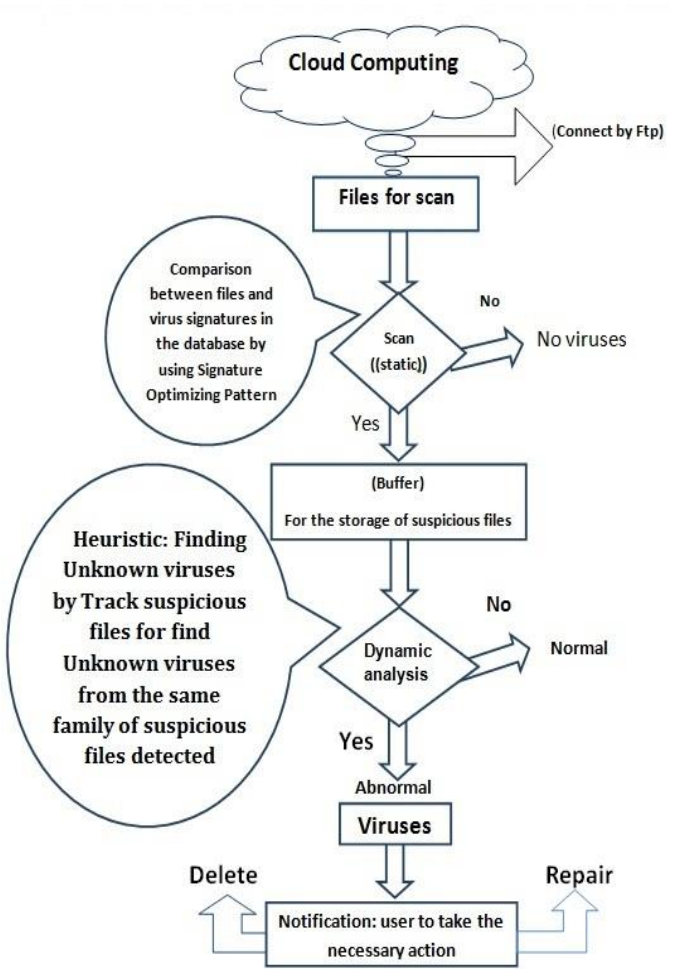


Fig. 3. Simple outline of processes used in proposed architecture

For experimental work, we used hosting services "000webhost.com" as cloud service for uploading the database and execute our system on it, show that in Figure 4

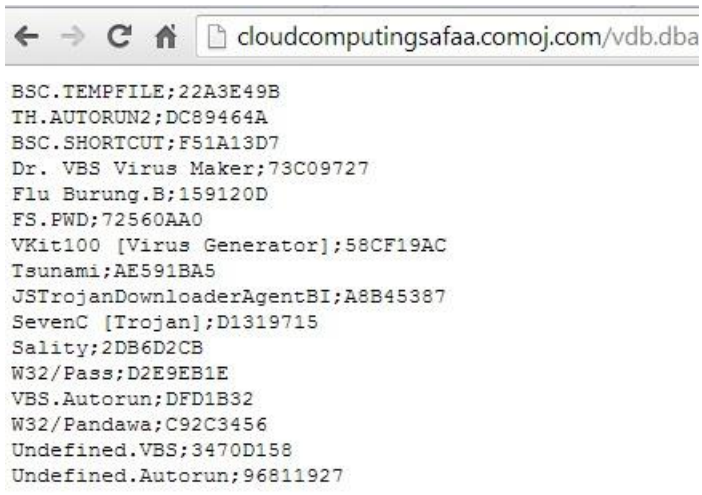


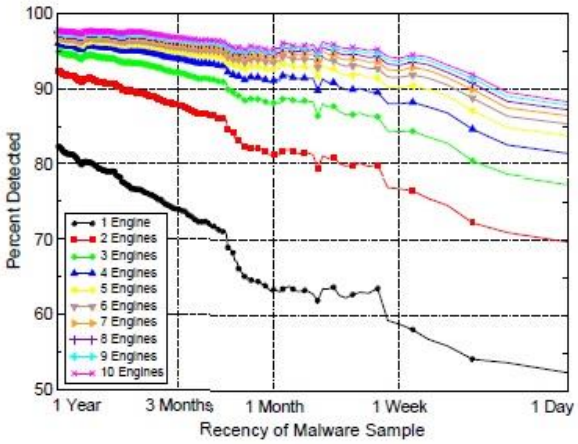
Fig. 4. Snapshot for virus signatures in our cloud environment

Oberheide ET. Al. [9] showed the detection rate with N-Version of Protection in the Network Cloud. N-version protection is closely related to our approach, a paradigm in which multiple implementations of critical software are written by independent parties to increase the reliability of software by reducing the probability of concurrent failures [15]. Traditionally, N-version programming has been applied to systems requiring high availability such as distributed file systems [16]. N-version programming has also been applied to the security realm to detect implementation faults in web services that may be exploited by an attacker [17].

Moreover, A handful of online services has recently been constructed that implement N-version detection techniques. For example, there are online web services for malware submission and analysis [18, 19, and 20]. However, these services are designed for the occasional manual upload of a virus sample, rather than the automated and real-time protection of end hosts, which results in vastly different architectural decisions and performance characteristics.

Engines	3 Months	1 Month	1 Week
1	73.9%	63.1%	59.6%
2	87.7%	81.0%	77.6%
3	92.0%	87.8%	84.8%
4	93.8%	90.9%	88.4%
5	94.8%	92.4%	90.5%
6	95.4%	93.4%	91.8%
7	95.9%	94.0%	92.8%
8	96.2%	94.5%	93.5%
9	96.5%	94.8%	94.0%
10	96.7%	95.0%	94.4%

(a)



(b)

Fig. 5. Detection rate for ten popular antivirus products as a function of the age of the malware samples [9]

In figure 5 (a) demonstrates how the use of multiple heterogeneous engines allows cloud to significantly improve the aggregate detection rate. While in figure 5 (b) shows the detection rate over malware samples ranging from one day old to one year old. The graph shows how using ten engines can increase the detection rate for the entire year-long AML dataset as high as 98%.

The proposed system uses multiple engines detection, and then uploads this process and database to cloud storage by using FTP (File Transfer Protocol), for dealing easily with large files and huge databases of viruses.

Consequently, our results Increase rates in detection rates up to 98% with an increase in the speed of detection and time spent and easy dealing with large files back up.

As Li [6] improved the detection by using signature matching optimization policy for anti-virus programs, we also, combine heuristic and signature matching optimization for detecting known viruses and unknown viruses not for known signature only, this increased detection rate as well as the development of system databases And exploit Time and effort in the detection of signatures in the full library.

IV. CONTRIBUTIONS AND REMARKS

Contributions

Our approach of moving the detection of malicious software into the cloud is aligned with a strong trend toward moving services from end host and monolithic servers into the cloud. For example, in-network email [21, 22] and HTTP [23] filtering systems are already popular and are used to provide an additional layer of security for enterprise networks. In addition, there have been several attempts to provide cloud services as overlay networks [24]; the main relevant contributions of our approach are the following:

The first main contribution of this thesis is the design and development of malware detection system in cloud, a system for providing anti-virus scanning for desktop computers. Cloud – malware detection (CMD) combines a set of pre-existing, third-party scanning services and offload the scanning of les from the host computer to these services. The evaluation of CMD found that performance of the system was highly dependent on the le system activity while the system was active, but that there were specific instances where the system performed well. The findings from this thesis can help to address the performance concerns involved in cloud-based malware scanning. This could result in a system that would be capable of performing nearly transparent anti-malware protection from the cloud.

The second contribution of this thesis was an extension of the desktop version of N-version protection. The system was designed and developed for the Window operating system, and the evaluation of the system showed favorable performance, suggesting that cloud-based anti-malware scanning may be a very good fit for providing a level of security to computer devices.

Finally, our thesis includes a comprehensive examination and summary of the current body of academic research pertaining to cloud-based security for both desktop computers and mobile devices, as well as research regarding low-impact anti-malware techniques which might also be suitable for Detect malware in cloud computing.

A. Why cloud computing applied to C.M.D?

1) Reduction of costs – unlike on-site hosting the price of deploying applications in the cloud can be less due to lower hardware costs from a more effective use of physical resources.

2) Universal access - cloud computing can allow remotely located employees to access applications and work via the Internet.

3) Up-to-date software - a cloud provider will also be able to upgrade software keeping in mind feedback from previous software releases.

4) Choice of applications- This allows flexibility for cloud users to experiment and choose the best option for their needs. Cloud computing also allows a business to use, access and pay only for what they use, with a fast implementation time.

5) Potential to be greener and more economical - the average amount of energy needed for a computational action carried out in the cloud is far less than the median amount for an on-site deployment. This is because different organizations can share the same physical resources securely, leading to more efficient use of the shared resources.

6) Flexibility – cloud computing allows users to switch applications easily and rapidly, using the one that suits them needs best. However, migrating data between applications can be an issue.

The proposed system includes two types of protection built in remote-server protection; make sure that it has a backup system by File Transfer Protocol (FTP); FTP is normally used to transfer files between computers on a network. Cloud FTP enables files to be transferred to Storage Clouds, for transforming data and process to the cloud.

Consequently ,these processes saves latest malware protection in a local cache on your computer then send to cloud server, So that it protects your PC even when you aren't connected to the cloud.

Elsewhere, our system also features a Lightweight, Dispute of the famous anti-virus Products. Figure 6, shows the effect of cloud on size comparison between Cloud malware detection (CMD) and famous antivirus.

Thus, in today's antivirus programs, static analysis is used in combination with dynamic analysis. The idea behind this combined approach is to emulate the execution of an application in a secure virtual environment; the following figure shows the detection rates for viruses of this system and Interface scan.

In view of different detection, methods must be combined to determine whether a file is secure to open, access, or execute. Several variables may impact this process, to be more powerful and more-safe at malware Known and unknown to continuous update of the database of viruses and automatically.

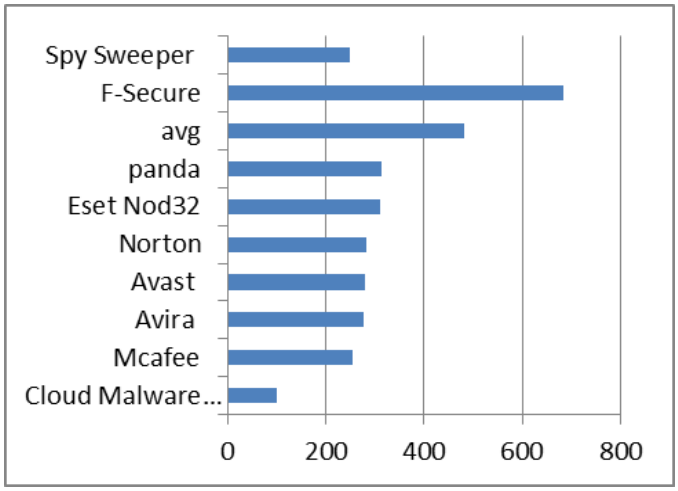


Fig. 6. Size of each installed anti-virus software - has been download updates after installation to being included in the results.

V. Conclusions and Future Work

To conclude, we have proposed a system for combined malware detection systems and cloud computing environments, all running binaries and malware are intercepted by submitting to one or more analysis engines, a complete check against a signature database to detect yet unknown exploits or malware. We will suggest increasing in the dependence of cloud computing as consumers increasingly move to cloud computing platforms for their computing needs. In this paper, we reviewed previous work on malware detection, both conventional and in the presence of storage in order to determine the best approach for detection in the cloud. We also argue the benefits of distributing detection throughout the cloud and present a new approach to coordinate detection across the cloud.

In the proposed system, we have used traditional detection techniques (optimizing pattern) as per static signatures and dynamic detection technology (heuristic). Then, we have chosen for safer system methods as well as speed and modern to rival existing anti-virus.

The proposal of this work is to find the best solutions to the problems of anti-viruses and improve performance and find possible alternatives for a better working environment without problems with high efficiency and flexibility.

We used the optimal traditional methods and modern to detect viruses, for unknown and already detected viruses through the signatures and the Heuristic.

Future work in this field will focus on the development of detection systems based on memory introspection and heuristic or statistical detection, as opposed to signature-based detection.

References

[1] Microsoft, “Microsoft security intelligence report”, [online]:http://www.microsoft.com/technet/security/default.mspx, July December 2006.

[2] Dropbox, Inc., dropbox.com webpage, [Online]: https://www.dropbox.Com/ (accessed 13/04/12).

[3] C. Grace. “Understanding intrusion-detection systems" [J], PC Network Advisor, vol. 122, pp. 11-15, 2000.

[4] S. Subashini, V. Kavitha s.l “A survey of security issues in service delivery models of cloud computing.” .Science Direct, Journal of Network and Computer Applications, pp. (1-11) January (2011).

[5] Shirlei Aparecida de Chaves, Rafael Brundo Uriarte and Carlos Becker Westphall “Toward an Architecture for Monitoring Private Clouds." S.l. IEEE December (2011).

[6] Bo Li, Eul Gyu I'm “ A signature matching optimization policy for anti-virus programs” Electronics and Computer Engineering, Hanyang University, Seoul, Korea. © IEEE 2011

[7] Chen, Z. & Yoon, J. “IT auditing to assure a secure cloud computing”, (2010). [Online]: http://doi.ieeecomputersociety.orgezproxy.umuc.edu/10.1109/SERVICES.2010.118.

[8] J. Oberheide, E. Cooke, and F. Jahanian “CloudAV: N-Version Antivirus in the Network Cloud”, In Proceedings of the 17th USENIX Security Symposium (Security'08). San Jose, CA, 2008.

[9] Jon Oberheide, Evan Cooke and Farnam Jahanian “Cloud N-Version Antivirus in the Network Cloud”,Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI 48109 (2007).

[10] Matthias Schmidt, Lars Baumg Artner, Pablo Graubner, David Bock and Bernd Freisleben “Malware Detection and Kernel Rootkit Prevention in Cloud Computing Environments." University of Marburg, Germany (2011).

[11] K. Murad, S. Shirazi, Y. Zikria, and I. Nassar, “Evading Virus Detection Using Code Obfuscation” in Future Generation Information Technology, vol. 6485 of Lecture Notes in Computer Science, pp. 394–401, Springer Berlin , Heidelberg, 2010.

[12] Scott Treadwell, Mian Zhou “A Heuristic Approach for Detection of Obfuscated Malware”, Bank of America, 1201 Main St, Dallas, TX 75202, © IEEE 2009.

[13] Carlin, S., & Curran, K. “Cloud computing security”, International Journal of Ambient Computing and Intelligence.

[14] “Heuristic analysis in Kaspersky Internet Security” [Online]: http://support.kaspersky.com , ID: 8936 , 2013 Mar 01 2013

[15] Algirdas Avizienis, “The n-version approach to fault-tolerant software”, IEEE Transactions on Software Engineering, 1985.

[16] Rodrigo Rodrigues, Miguel Castro, and Barbara Liskov. Base, “using abstraction to improve fault tolerance”, In Proceedings of the eighteenth ACM symposium on Operating systems principles, New York, NY, USA, 2001.

[17] Lajos Nagy, Richard Ford, and William Allen, “N-version programming for the detection of zero-day exploits”, In IEEE Topical Conference on Cybersecurity, Daytona Beach, Florida, USA, 2006.

[18] Carsten Willems and Thorsten Holz. Cwsandbox.[Online]: http://www.cwsandbox.org/, 2007.

[19] Hispasec Sistemas. “Virus total”, [Online]: http://virustotal.com, 2004.

[20] Norman Solutions. Norman sandbox whitepaper. http://download.norman.no/whitepapers/whitepaper_Norman_SandBox.pdf, 2003.

[21] Barracuda Networks. “Barracuda spam firewall”,[Online]: http://www.barracudanetworks.com, 2007.

[22] Cloudmark, “Cloudmark authority anti-virus”,[Online] :http://www.cloudmark.com, 2007.

[23] Alexander Moshchuk, Tanya Bragin, Damien Deville, Steven D. Gribble, and Henry M. Levy, “Spyproxy: Execution-based detection of malicious web content”, In Proceedings of the 16th USENIX Security Symposium, August 2007.

[24] Stelios Sidiroglou, Angelos Stavrou, and Angelos D. Keromytis, “Mediated overlay services (moses): Network security as a composable service”, In Proceedings of the IEEE Sarnoff Symposium, Princeton, NJ, USA, 2007.

EEG Mouse:A Machine Learning-Based Brain Computer Interface

Mohammad H. Alomari, Ayman AbuBaker, Aiman Turani, Ali M. Baniyounes, Adnan Manasreh
Electrical and Computer Engineering Department, Applied Science University
P.O. Box 166, Amman 11931 Jordan

***Abstract*—The main idea of the current work is to use a wireless Electroencephalography (EEG) headset as a remote control for the mouse cursor of a personal computer. The proposed system uses EEG signals as a communication link between brains and computers. Signal records obtained from the PhysioNet EEG dataset were analyzed using the Coif lets wavelets and many features were extracted using different amplitude estimators for the wavelet coefficients. The extracted features were inputted into machine learning algorithms to generate the decision rules required for our application. The suggested real time implementation of the system was tested and very good performance was achieved. This system could be helpful for disabled people as they can control computer applications via the imagination of fists and feet movements in addition to closing eyes for a short period of time.**

Keywords—EEG; BCI; Data Mining; Machine Learning; SVMs; NNs; DWT; Feature Extraction

I. INTRODUCTION

Brain-Computer Interface (BCI) is a device that enables the use of the brain's neural activity to communicate with others or to control machines, artificial limbs, or robots without direct physical movements [1-4]. As computerized systems are becoming one of the main tools for making people's lives easier and with the ongoing growth in the BCI field, it is becoming more important to understand brain waves and analyze EEG signals. Electroencephalography (EEG) is the process of measuring the brain's neural activity as electrical voltage fluctuations along the scalp as a result of the current flows in brain's neurons [5]. The brain's electrical activity is monitored and recorded, in typical EEG tests, using electrodes that are fixed on the scalp [6]. BCI captures EEG signals in conjunction with a specific user activity then uses different signal processing algorithms to translate these records into control commands for different machine and computer applications [7].

BCI was known for its popular use in helping disabled individuals by providing a new channel of communication with the external environment and offering a feasible tool to control artificial limbs [8]. A variety of BCI applications were described in[9] including the control of devices using the translation of thoughts into commands in video games and personal computers. BCI is a highly interdisciplinary research topic that combines medicine, neurology, psychology, rehabilitation engineering, Human-Computer Interaction (HCI), signal processing and machine learning [10].

In our previous research [11-13] we proposed many systems that could efficiently discriminate between executed (or imagined) left and right fist (or feet) movements. In this work, we integrated these systems into one hybrid application that is based on the imagined fists and feet movements.

II. LITERATURE REVIEW

The translation approach used to transform EEG signal patterns into machine commands reflects the strength of BCI applications. In [14], the authors recorded EEG signals for three subjects while imagining either right or left hand movement based on a visual cue stimulus. They were able to classify EEG signals into right and left hand movements using a neural network classifier with an accuracy of 80% and concluded that this accuracy did not improve with increasing number of sessions.

The authors of[15] used features produced by Motor Imagery (MI) to control a robot arm. Features such as the band power in specific frequency bands (alpha: 8-12Hz and beta: 13-30Hz) were mapped into right and left limb movements. In addition, they used similar features with MI, which are the Event Related Resynchronization and Synchronization (ERD/ERS) comparing the signal's energy in specific frequency bands with respect to the mentally relaxed state.

The combination of ERD/ERS and Movement-Related Cortical Potentials (MRCP) was proven to improve the classification of EEG signals as this offers an independent and complimentary information [13, 16]. The authors of [17]presented an approach for the classification of single trial MRCP using a discrete dyadic wavelet transform and Support Vector Machines (SVMs) and they provided a promising classification performance.

A single trial right/left hand movement classification is reported in [18]. The authors analyzed both executed and imagined hand movement EEG signals and created a feature vector consisting of the ERD/ERS patterns of the mu and beta rhythms and the coefficients of the autoregressive model. Artificial Neural Networks (ANNs) is applied to two kinds of testing datasets and an average recognition rate of 93% is achieved.

A three-class BCI system was presented in [19] for the translation of imagined left/right hands and foot movements into commands that operates a wheelchair. This work used many spatial patterns of ERD on mu rhythms along the sensory-motor cortex and the resulting classification accuracy

for online and offline tests was 79.48% and 85.00%, respectively. The authors of [20] proposed an EEG-based BCI system that controls hand prosthesis of paralyzed people by movement thoughts of left and right hands. They reported an accuracy of about 90%.

In [21], a hybrid BCI control strategy is presented. The authors expanded the control functions of a P300 potential based BCI for virtual devices and MI related sensorimotor rhythms to navigate in a virtual environment. Imagined left/right hand movements were translated into movement commands in a virtual apartment and an extremely high testing accuracy results were reached.The Daubechies, Coiflet and Symmlet wavelet families were applied in[22]to a dataset of MI to extract features and describe right and left hand movement imagery. The authors reported that the use of Linear Discriminate Analysis (LDA) and Multilayer Perception (MLP) Neural Networks (NNs) provided good classification results and that LDA classifier achieved higher classification results of up to 88% for different Symmlet wavelets. The authors of[23]used the discrete wavelet transform (DWT) to create inputs for a NNs classifier and the authors reported a very high classification accuracy of 99.87% for the recognition of some mental tasks.

III. The Proposed System

The main idea of the current work is to use a wireless EEG headset such as the one designed by NeuroSky[24] as a remote control for the mouse cursor of personal computers and the computer applications. As depicted in Fig. 1, the captured EEG signals have to be pre-processed to filter out the unwanted content and then the content of interest has to be represented using some features that can be inputted into machine learning algorithms. The outcome of this process is a collection of decision rules that can be translated, as required, into PC commands.

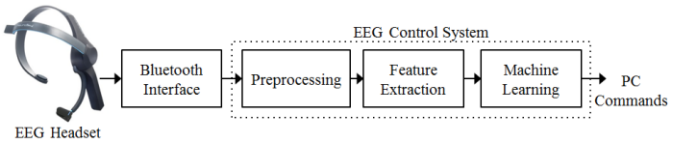


Fig. 1. A block diagram for the suggested system

A. Eeg Data

The PhysioNetEEG dataset [25] is used in this work. It consists of more than 1500 one or two minutes-duration EEG records obtained from 109 healthy subjects. Subjects were asked to execute and imagine different tasks while 64 channels of EEG signals were recorded from the electrodes that were fitted along the scalp.

In the records of the dataset that are related to the current research, each subject performed the following tasks:

- One-minute baseline run with eyes open.
- One-minute baseline run with eyes closed.
- Three two-minutes experimental runs of imagining moving theright or left fists while the left or right side of a computer screen is showing a target.
- Three two-minute experimental runs of imagining moving both fists or both feet while the top or bottom side of a computer screen is showing a target.

The obtained EEG signals were recorded according to the international 10-20 system as seen in Fig. 2.

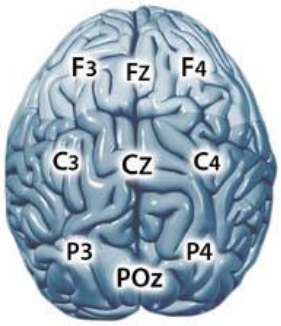


Fig. 2. Electrode Positions for the C3, Cz, and C4 channels

For this work, we created a subset for 100 subjects including 8 runs per subject.

B. Preprocessing

Only channels C3, C4, and Cz were used in our work for two reasons:(1) It is reported in [6] that most EEG channels represent redundant information and (2), it was concludedin[26, 27] that the neural activity that is mostly correlated to the fists movements is almost exclusively contained within these channels as depicted in Fig. 2.

The authors of [28] showed thatEEG signals are noisy and non-stationary signals that have to be filtered to get rid of the unnecessary content. Hence, the channels C3, C4, and Cz were filtered, using a band-pass filter (0.5-50 Hz),for the purpose of removing the DC shifts and minimizing the presence of filtering artifacts at epoch boundaries.

In [29], it was stated that EEG signals are usually masked by physiological artifacts that produce huge amounts of useless data. Eye and muscle movements could be good examples of these artifacts that constitute a challenge in the field of BCI research. The Automatic Artifact Removal (AAR) toolbox [30] was used to process our EEG subset.

A MATLAB script was written to analyze the filtered EEG signals and it was found that a subject imagines opening and closing a fist (or both fists/feet) and keeps doing this for 4.1 seconds then he takes a rest for the duration of 4.2 seconds. This means that each two-minute EEG run includes 15 events that are separated by a short neutral period where the subject relaxes. As the Physionet dataset was sampled at 160 samples per second, each vector includes 656 samples of the original recorded EEG signal. And because we used the available records for 100 subjects, our subset included 18000 vectors representing imagined left fist, right fist, both fists, and both feet movements. An additional 1500 vectors were extracted from the one-minute baseline run (with eyes open) and another 1500 vectors from the one-minute baseline run with eyes closed. So, the total number of extracted samples (events) was 12000 samples.

IV. Feature Extraction

A. The Discrete Wavelet Transform

The Wavelet transform analysis was used in a wide range of research topics within the field of signal processing. Based on a multi-resolutions process, the wavelet properties of a scalable window allow pinpointing signal components. These properties of dilation and translation enable the extraction of all components for every position by creating different scales and shifted functions (in time domain) of a signal [31, 32]. As a result, wavelet finer and large scaling, permit all information of the signal (the big picture), while small scales shows signal details by zooming into the signal components.

For discrete data, such as the datasets used in the current work, the Discrete Wavelet Transform (DWT) is better fit for analysis. It was shown in [33] that a suitable wavelet function must be used to optimize the analysis performance. A large selection of DWT mother wavelets, such as the Daubechies, Symmlet, and Coif let, is available to be used in our work [22]. But the Coif let(Coif) family of wavelet functions provided the best classification performance in our previous work [11]. So, we decided to calculate the Coif lets wavelets Coif1-Coif5 in this work.

As shown in Fig. 3, the main purpose of the DWT is to decompose the recorded EEG signal into multi-resolution subsets of coefficients: a detailed coefficient subset(cD_i) and an approximation coefficient subset (cA_i) at thelevel*i*.So, at the first decomposition level we obtain cD_1 and cA_1 then the first approximation cA_1 can be transformed into cD_2 and cA_2 at the second level, and so on. For our experiments, the decomposition level was set to generate four level details.

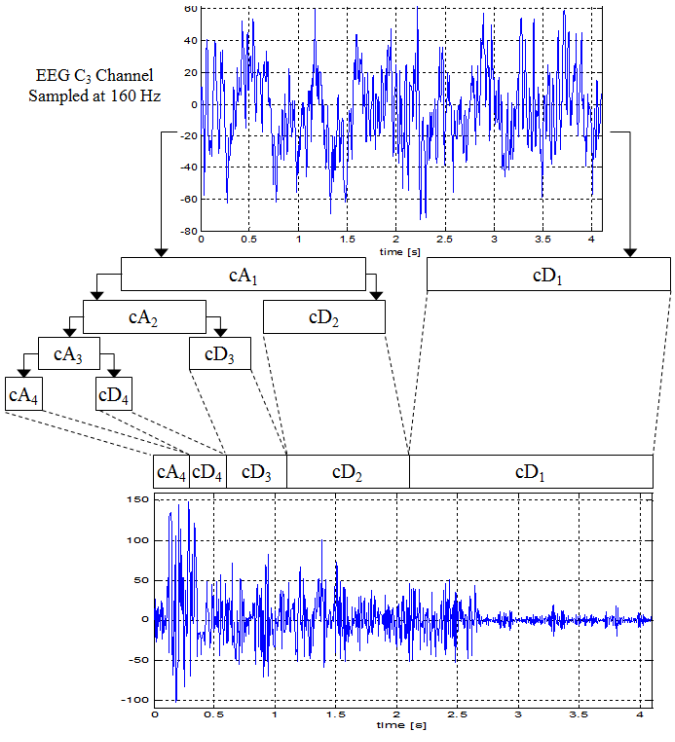


Fig. 3. The multi-resolution decomposition of a sample EEG signal.

All EEG signals in the subset were sampled at 160Hz. So, the wavelet transformation of each record at four levels results in four details: cD_1(40-80Hz), cD_2(20-40Hz), cD_3(10-20Hz), and cD_4(5-10Hz), and a single approximationcA_4(0-5Hz). As explained in [11], the details cD_2, $cD_3$and cD_4 provided proper representation for the activities of interest. So, we decided to extract the vectors of features from these details only.

B. Amplitude Estimators

Many amplitude estimators for neurological activities were defined mathematically in [34] and some of them were selected based on our previous results obtained in [11].

If we assume that the n^{th} sample of a wavelet decomposed detail at level i is $D_i(n)$, then the following features can be defined:

- Root Mean Square (RMS)

$$RMS_i = \sqrt{\frac{1}{N}\sum_{n=1}^{N} D_i^2(n)} \quad (1)$$

- Mean Absolute Value (MAV)

$$MAV_i = \frac{1}{N}\sum_{n=1}^{N} |D_i(n)| \quad (2)$$

- Integrated EEG (IEEG)

$$IEEG_i = \sum_{n=1}^{N} |D_i(n)| \quad (3)$$

- Simple Square Integral (SSI)

$$SSI_i = \sum_{n=1}^{N} |D_i(n)|^2 \quad (4)$$

- Variance of EEG (VAR)

$$VAR_i = \frac{1}{N-1}\sum_{n=1}^{N} D_i^2(n) \quad (5)$$

- Average Amplitude Change (AAC)

$$AAC_i = \frac{1}{N}\sum_{n=1}^{N} |D_i(n+1) - D_i(n)| \quad (6)$$

C. Feature Vectors

In our experiments, we applied the Coifletswavelets Coif1 to Coif5 for each one of the channels C3, C4, and Cz of an EEG record. This process was repeated for each event in our dataset of 12000 vectors.

Then, all estimators were calculated using (1) through (6) for the details cD_2, $cD_3$and cD_4 of each instance.At the end of these calculations, 9 features of each estimator (3 channels ×3 details) were generated for each Coiflets wavelet. These features were numerically represented in a format that is suitable for use with SVMs and NNs algorithms [35, 36] as described in the next section.

V. Machine Learning

SVMs and NNs learning algorithms were used in [13, 14, 22, 23, 37] and provided excellent classification performances. A detailed description of SVMs and NNs can be found in[36]. The MATLAB NN toolbox was used for all the training and testing experiments. The training experiments were handled with the aid of the back-propagation learning algorithm [38].

SVM experiments were carried out using the "MySVM" software[39]. SVM can be performed with different kernels and most of them were reported to provide similar results for similar applications [6]. So, the Anova-Kernel SVM was used in this work.

As shown on Fig. 4, NNs and SVMs classifiers were created with 9 inputs, representing features of one estimator. The SVM classifier has one output node representing the target function: closed eyes/opened eyes. The NN classifier has one output node that has five possible classes: opened eyes, left fist, right fist, both fists, and both feet. Both classifiers were integrated such that the NN classifier is only enabled when the eyes are open.

In SVM, each of the degree and gamma parameters were varied from 1 to 10 and the number of hidden layers for the neural network was varied from 1 to 20.

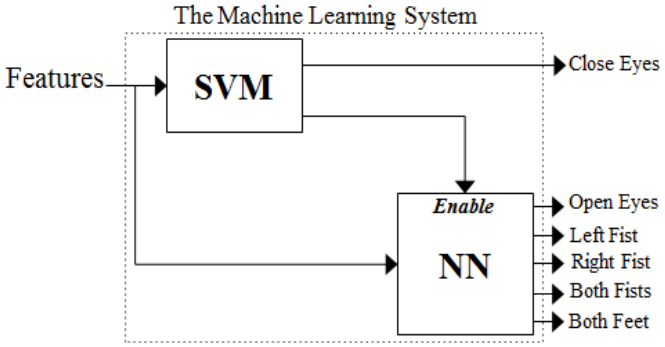


Fig. 4. The Hybrid Machine Learning System

At each specific number of hidden layers (or a specific degree-gamma pair), 80% of the samples (9600 events) were randomly selected and used for training and the remaining 20%for testing. This process was repeated 10 times, and in each time the datasets were randomly mixed. For each specific configuration, the average accuracy was calculated for the ten training-testing pairs.

A huge number of training and testing experiments were carried out. Table I and Table II list the best average accuracies with their associated classifier configurations. It can be noted that the use of a SVMs classifier of gamma = 9 and degree = 6 with inputs that were generated by a Coif4 wavelet and MAV features provided the optimum classification performance of an accuracy of 74.97%. In addition, a NNs classifier of 15 hidden layers with inputs that were generated by a Coif2 wavelet and IEEG features provided an accuracy of 71.6%. These are very promising results as they were obtained while most of the available data are for imagined movements.

TABLE I. Optimum Classification Results Achieved Using Different Coifletswavelets With Svms.

Features	MAV			RMS			AAC			IEEG			SSI			VAR		
Coiflets wavelet	***gam***	***deg***	***AvgAcc***	***gam***	***deg***	***AvgAcc***	***gam***	***deg***	***AvgAcc***	***gam***	***deg***	***AvgAcc***	***gam***	***deg***	***AvgAcc***	***gam***	***deg***	***AvgAcc***
Coif1	3	6	0.7011	3	5	0.6911	9	7	0.6821	5	8	0.6930	4	6	0.6183	8	8	0.6011
Coif2	8	5	0.6903	9	2	0.6857	3	6	0.6532	6	5	0.6814	3	5	0.6634	5	2	0.6122
Coif3	3	4	0.7152	6	2	0.7033	6	9	0.6642	8	7	0.6598	3	7	0.6120	6	5	0.5984
Coif4	9	6	0.7497	8	7	0.7112	8	3	0.6803	3	4	0.6786	8	3	0.6045	4	6	0.6103
Coif5	9	5	0.7325	4	5	0.7058	2	2	0.6792	6	3	0.6133	8	4	0.6143	5	7	0.6002

TABLE II. Optimum Classification Results Achieved Using Different Coifletswavelets With Nns.

Features	MAV		RMS		AAC		IEEG		SSI		VAR	
Coiflets wavelet	***HL***	***AvgAcc***	***HL***	***AvgAcc***	***HL***	***AvgAcc***	***HL***	***AvgAcc***	***HL***	***AvgAcc***	***HL***	***AvgAcc***
Coif1	16	0.6166	14	0.6186	17	0.5801	19	0.6612	19	0.6247	13	0.5781
Coif2	20	0.6470	19	0.6430	13	0.5821	15	0.7160	19	0.5862	12	0.5801
Coif3	16	0.5882	19	0.6349	18	0.5923	20	0.6207	13	0.5578	18	0.6491
Coif4	16	0.5984	16	0.6186	19	0.6045	18	0.6065	9	0.5538	15	0.5538
Coif5	11	0.6247	18	0.6045	20	0.5984	19	0.6227	13	0.5335	17	0.5396

VI. Real Time Implementation

A simple software interface was designed as show in Fig. 5. This software reads streams of EEG signals from a test EEG record or from an EEG mouse (if available).

The system extracts the features needed for the SVM and NN decision rules and provides near-real time actions. The default configurations of this system are to translate the "closing eyes for 2s" activity into a mouse click and the

imagined left/right fists, both fists, and both feet movements into computer cursor movements as seen in Fig. 6. These configurations can be reprogrammed to run different computer applications instead of simple cursor movements.

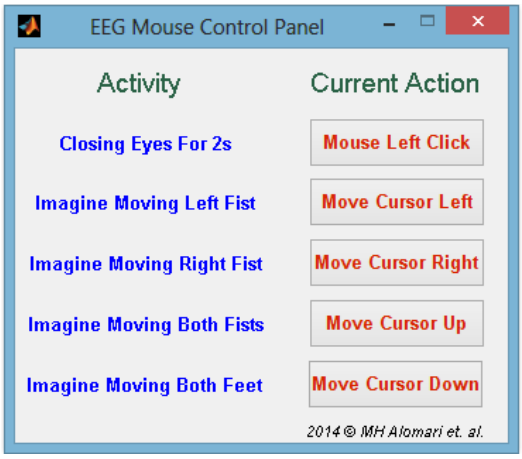


Fig. 5. EEG Mouse Control Panel

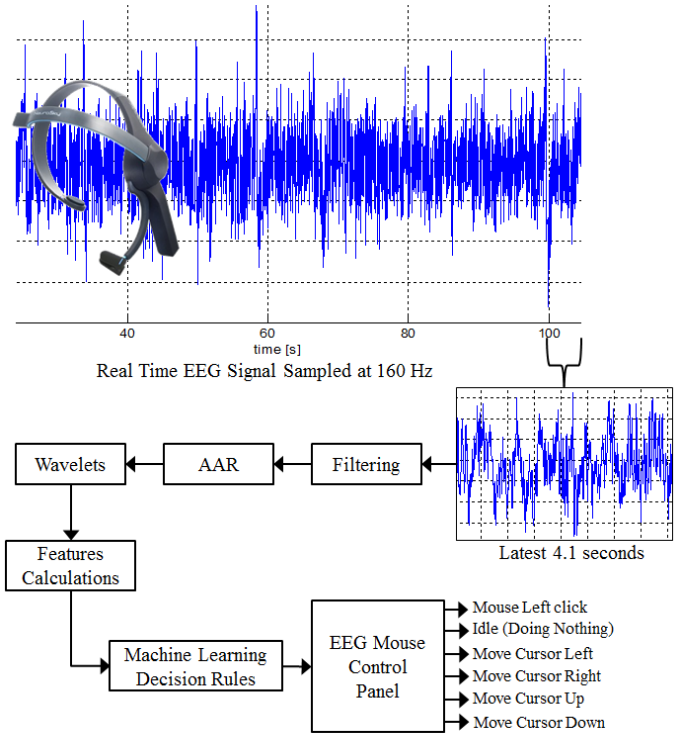


Fig. 6. The Suggested Real-Time Implementation of the System

VII. Conclusions

The objective of this work was to enable the use of the available commercial EEG headsets as a remote control for computer applications. Disabled people may use this system as a channel of communication with computers and they can provide some simple computer commands by imagination. Signal records obtained from the PhysioNet EEG dataset were analyzed using the Coiflets wavelets and machine learning algorithms and promising classification performances were obtained.

References

[1] J. P. Donoghue, "Connecting cortex to machines: recent advances in brain interfaces," Nature Neuroscience Supplement, vol. 5, pp. 1085–1088, 2002.

[2] S. Levine, J. Huggins, S. BeMent, R. Kushwaha, L. Schuh, E. Passaro, M. Rohde, and D. Ross, "Identification of electrocorticogram patterns as the basis for a direct brain interface," Journal of Clinical Neurophysiology, vol. 16, pp. 439-447, 1999.

[3] A. Vallabhaneni, T. Wang, and B. He, "Brain—Computer Interface," in Neural Engineering, B. He, Ed.: Springer US, 2005, pp. 85-121.

[4] [4] J. Wolpaw, N. Birbaumer, D. McFarland, G. Pfurtscheller, and T. Vaughan, "Brain-computer interfaces for communication and control," Clinical Neurophysiology, vol. 113, pp. 767-791, 2002.

[5] E. Niedermeyer and F. H. L. da Silva, Electroencephalography: Basic Principles, Clinical Applications, and Related Fields: Lippincott Williams & Wilkins, 2005.

[6] J. Sleight, P. Pillai, and S. Mohan, "Classification of Executed and Imagined Motor Movement EEG Signals," Ann Arbor: University of Michigan, 2009, pp. 1-10.

[7] B. Graimann, G. Pfurtscheller, and B. Allison, "Brain-Computer Interfaces: A Gentle Introduction," in Brain-Computer Interfaces: Springer Berlin Heidelberg, 2010, pp. 1-27.

[8] A. E. Selim, M. A. Wahed, and Y. M. Kadah, "Machine Learning Methodologies in Brain-Computer Interface Systems," in Biomedical Engineering Conference, 2008. CIBEC 2008. Cairo International, 2008, pp. 1-5.

[9] E. Grabianowski, "How Brain-computer Interfaces Work," HowStuffWorks, Inc, 2007.

[10] M. Smith, G. Salvendy, K. R. M√¨ller, M. Krauledat, G. Dornhege, G. Curio, and B. Blankertz, "Machine Learning and Applications for Brain-Computer Interfacing," in Human Interface and the Management of Information. Methods, Techniques and Tools in Information Design. vol. 4557: Springer Berlin Heidelberg, 2007, pp. 705-714.

[11] M. H. Alomari, E. A. Awada, A. Samaha, and K. Alkamha, "Wavelet-Based Feature Extraction for the Analysis of EEG Signals Associated with Imagined Fists and Feet Movements," Computer and Information Science, vol. 7, pp. 17-27, 2014.

[12] M. H. Alomari, E. A. Awada, and O. Younis, "Subject-Independent EEG-Based Discrimination Between Imagined and Executed, Right and Left Fists Movements," European Journal of Scientific Research, vol. 118, pp. 364-373, 2014.

[13] M. H. Alomari, A. Samaha, and K. AlKamha, "Automated Classification of L/R Hand Movement EEG Signals using Advanced Feature Extraction and Machine Learning," International Journal of Advanced Computer Science and Applications, vol. 4, pp. 207-212, 2013.

[14] G. Pfurtscheller, C. Neuper, D. Flotzinger, and M. Pregenzer, "EEG-based discrimination between imagination of right and left hand movement," Electroencephalography and Clinical Neurophysiology, vol. 103, pp. 642-651, 1997.

[15] F. Sepulveda, "Brain-Actuated Control of Robot Navigation," in Advances in Robot Navigation, A. Barrera, Ed.: InTech, 2011.

[16] A.-K. Mohamed, "Towards improved EEG interpretation in a sensorimotor BCI for the control of a prosthetic or orthotic hand," in Faculty of Engineering. vol. Thesis: Master of Science in Engineering Johannesburg: University of Witwatersrand, 2011, p. http://hdl.handle.net/10539/10546.

[17] D. Farina, O. F. d. Nascimento, M.-F. Lucas, and C. Doncarli, "Optimization of wavelets for classification of movement-related cortical potentials generated by variation of force-related parameters," Journal of Neuroscience Methods, vol. 162, pp. 357-363, 5/15/ 2007.

[18] J. A. Kim, D. U. Hwang, S. Y. Cho, and S. K. Han, "Single trial discrimination between right and left hand movement with EEG signal," in Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Cancun, Mexico, 2003, pp. 3321-3324 Vol.4.

[19] Y. Wang, B. Hong, X. Gao, and S. Gao, "Implementation of a Brain-Computer Interface Based on Three States of Motor Imagery," in 29th

Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007. EMBS 2007., 2007, pp. 5059-5062.

[20] C. Guger, W. Harkam, C. Hertnaes, and G. Pfurtscheller, "Prosthetic Control by an EEG-based Brain- Computer Interface (BCI)," in AAATE 5th European Conference for the Advancement of Assistive Technology, Düsseldorf, Germany, 1999.

[21] Y. Su, Y. Qi, J.-x. Luo, B. Wu, F. Yang, Y. Li, Y.-t. Zhuang, X.-x. Zheng, and W.-d. Chen, "A hybrid brain-computer interface control strategy in a virtual environment," Journal of Zhejiang University SCIENCE C, vol. 12, pp. 351-361, 2011.

[22] I. Homri, S. Yacoub, and N. Ellouze, "Optimal segments selection for EEG classification," in 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), Sousse, Tunisia, 2012, pp. 817-821.

[23] M. Tolić and F. Jović, "Classification of Wavelet Transformed EEG Signals with Neural Network for Imagined Mental and Motor Tasks," International Journal of Fundamental and Applied Kinesiology, vol. 45, pp. 130-138, 2013.

[24] NeuroSky, "MindWave Mobile: MyndPlay Bundle," in EEG Biosensor Solutions: http://neurosky.com/products-markets/eeg-biosensors, 2014.

[25] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," Circulation, vol. 101, pp. e215-e220, June 13, 2000 2000.

[26] L. Deecke, H. Weinberg, and P. Brickett, "Magnetic fields of the human brain accompanying voluntary movements: Bereitschaftsmagnetfeld," Experimental Brain Research, vol. 48, pp. 144-148, 1982.

[27] C. Neuper and G. Pfurtscheller, "Evidence for distinct beta resonance frequencies in human EEG related to specific sensorimotor cortical areas," Clinical Neurophysiology, vol. 112, pp. 2084-2097, 2001.

[28] R. Romo-Vazquez, R. Ranta, V. Louis-Dorr, and D. Maquin, "EEG Ocular Artefacts and Noise Removal," in 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2007. , Lyon, 2007, pp. 5445-5448.

[29] G. Bartels, S. Li-Chen, and L. Bao-Liang, "Automatic artifact removal from EEG - a mixed approach based on double blind source separation and support vector machine," in 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2010, pp. 5383-5386.

[30] G. Gómez-Herrero, "Automatic Artifact Removal (AAR) toolbox for MATLAB," in Transform methods for Electroencephalography (EEG): http://kasku.org/projects/eeg/aar.htm, 2008.

[31] S. Tuntisak and S. Premrudeepreechacharn, "Harmonic Detection in Distribution Systems Using Wavelet Transform and Support Vector Machine," in 2007 IEEE Lausanne Power Tech, Lausanne, 2007, pp. 1540-1545.

[32] L. Qingyang and S. Zhe, "Method of Harmonic Detection Based On the Wavelet Transform," in International Conference on Information and Computer Applications (ICICA), Hong Kong, 2012, pp. 213-217.

[33] A. Phinyomark, C. Limsakul, and P. Phukpattaranont, "Optimal Wavelet Functions in Wavelet Denoising for Multifunction Myoelectric Control," ECTI Transactions on Electrical Eng., Electronics, and Communications, vol. 8, pp. 43-52, 2010.

[34] A. Phinyomark, F. Quaine, Y. Laurillau, S. Thongpanja, C. Limsakul, and P. Phukpattaranont, "EMG Amplitude Estimators Based on Probability Distribution for Muscle-Computer Interface," Fluctuation and Noise Letters, vol. 12, p. 1350016, 2013.

[35] M. Al-Omari, R. Qahwaji, T. Colak, and S. Ipson, "Machine learning-based investigation of the associations between cmes and filaments," Solar Physics, vol. 262, pp. 511-539, 2010.

[36] R. Qahwaji, T. Colak, M. Al-Omari, and S. Ipson, "Automated Prediction of CMEs Using Machine Learning of CME – Flare Associations," Sol. Phys, vol. 248, pp. 471-483 2008.

[37] P. A. Kharat and S. V. Dudul, "Daubechies Wavelet Neural Network Classifier for the Diagnosis of Epilepsy," Wseas Transactions on Biology and Biomedicine, vol. 9, pp. 103-113, 2012.

[38] S. E. Fahlmann and C. Lebiere, "The cascade-correlation learning architecture," in Advances in Neural Information Processing Systems 2 (NIPS-2) Denver, Colorado, 1989.

[39] S. Rüping, "mySVM-Manual ": University of Dortmund, Lehrstuhl Informatik 8, 2000.

Modeling and Forecasting the Number of Pilgrims Coming from Outside the Kingdom of Saudi Arabia Using Bayesian and Box-Jenkins Approaches

SAMEER M. SHAARAWY, ESAM A. KHAN, MAHMOUD A. ELGAMAL
The Custodian of the Two Holy Mosques Institute for Hajj and
Omra Research, Umm Al -Qura University
Makkah, Saudi Arabia

***Abstract*—Pilgrimage has received a great attention by the government of Saudi Arabia. Of special interest is the yearly series of the Number of Pilgrims coming from outside the kingdom (NPO) since it is one of the most important indicators in determining the planning mechanism for future hajj seasons. This study approaches the problems of identification, estimation, diagnostic checking and forecasting of the NPO series using Bayesian and Box - Jenkins approaches. The accuracy of Bayesian and Box- Jenkins techniques have been checked for forecasting the future observations and the results were very satisfactory. Moreover, it has been shown that Bayesian technique gives more accurate results than Box-Jenkins technique.**

Keywords—autoregressive processes, identification; estimation; diagnostic checking; forecasting; Jeffreys' prior; and posterior probability mass function.

I. INTRODUCTION

Pilgrimage (or hajj) of Moslems is one of the most important events all over the world. It is considered the largest human gathering in which pilgrims move together through a very limited space in a short period of time. This important event is repeated annually, and the number of pilgrims is increasing year after year. In addition, hajj is one of the main sources of gross national product (GNP) in Saudi Arabia. Therefore, it has received a great attention by the government of Saudi Arabia. Every year, the kingdom of Saudi Arabia spends a great deal of efforts and money to improve the hajj system, which includes security, economy, management of water and electrical resources, services and goods required by the vast number of pilgrims. However, without knowing the number of pilgrims in advance may make the process of improvement very difficult. Therefore, it is very important to have a mechanism for predicting and forecasting the number of pilgrims in order to determine the size and quality of expansions and maintenance needed in the two holy mosques in Makkah and Medina[1] and to avoid any mistakes or disasters that may occur. One of the main components of the total number of pilgrims is the Number of Pilgrims coming from outside the kingdom (NPO).

The first objective of this paper is to use the modern Bayesian approach to implement the identification, diagnostic checking and forecasting phases of NPO data. The foundation of the proposed Bayesian analysis is to use the pure autoregressive processes, denoted by AR (P) for short, to model and forecast the NPO data. There are three main reasons to use pure AR (P) processes to analyze our data. First, most data arise in real applications can be well presented by such processes. Second, the likelihood function OF pure AR (P) processes is analytically tractable because the white noise is a linear function of the model parameters, hence, as it will be seen in section 3, one may develop the exact posterior mass function of the model order. Third, it became clear to the authors, after a preliminary examination of the data, as it will be seen in section 4, that the pure AR (P) processes are appropriate to model and forecast the NPO data. The second objective of this paper is to use the well- known Box-Jenkins methodology to do a complete time series analysis of the NPO data. The final objective of this paper is to compare the accuracy of the results achieved by the Bayesian and Box-Jenkins approaches. he literature on time series is vast and can be found in many other areas other than statistics. Most of the literature is non- Bayesian and the reader is referred to Box-Jenkins (1970), Priestely (1981), Bowerman and O' Connell (1987), Tong (1990), Harvey (1993), Wei (2005) and Liu (2009). It is a fact that the methodology of Box-Jenkins is the most popular and prevailing traditional methodology to model and forecast time series. However, the Box-Jenkins methodology has serious disadvantages and drawbacks. Their identification technique is highly nonobjective and requires a very careful examination for the raw data and very good skills. In order to implement the identification stage, the time series analyst should be knowledgeable, well experienced and highly trained. In addition, he should have a large amount of data in order to identify an adequate model, see Chatfield (2004).

On the other hand, the Bayesian analysis of time series is still being developed and most of the Bayesian contributions have been occurred within the last three decades. Zellner (1971) introduced the subject for special autoregressive and econometric models. Newbold (1973) made an important contribution by his analysis of ARMA type transfer function

[1] **Makkah and Medina are the two well-known holy cities in Saudi Arabia where pilgrims should perform circumambulation in Makkah holy mosque: and most of them visit Medina mosque.*

models. Newbold's results were based on a t- approximation for the posterior analysis, as did the latter work of Zellner and Reynolds (1978). During this period, Bayesian forecasting was advanced by Chow (1975), who found the moments of the joint predictive distribution of future observations. One of the most important contributions of time series analysis was done by Monahan (1981), who used a numerical integration to implement the identification, estimation and forecasting phases of low order ARMA processes. This was the first Bayesian attempt to perform a numerical comprehensive time series analysis and was very valuable contribution. Shaarawy and Broemeling (1984) and Broemeling and Shaarawy (1988) have developed Bayesian techniques for identification, estimation, diagnostic checking and forecasting phases based on a t-approximation to the posterior distribution of the coefficients. Their first study has been extended later by Chen (1992) to bilinear model. Recently, Shaarawy and Ali (2003) have initiated a direct Bayesian technique to identify the orders of seasonal autoregressive processes. Their approach has been extended to the case of moving average processes by Shaarawy et al. (2007). The multivariate version of their direct approach has been introduced by Shaarawy and Ali (2008).

The Bayesian approach has several advantages when compared with Box-Jenkins approach, most obvious is pedagogical. It is much easier to learn the Bayesian methodology once one has mastered the inferential interpretation of Bayes' theorem. On the other hand, with traditional analysis, one must learn a large variety of sampling theory techniques. Second, the importance of Bayesian methods in economics, finance, engineering, education and other fields has increased rapidly over the last two decades. Third, the Bayesian methodology provides the time series analyst, in all areas of applications, with a formal and unifying way to incorporate the prior information in the analysis before seeing the data and this may lead to exact small sample results, see Broemeling and Shaarawy (1988). Fourth, our proposed Bayesian methodology does not assume stationarity.

The remainder of this paper is organized as follows: Section II presents autoregressive processes and processes and their basic characteristics. A complete Bayesian analysis for NPO is developed in section III. Section IV is devoted to model and forecast the NPO using the traditional method developed by Box and Jenkins (1970). Section V is dedicated to evaluate the forecast performance of Bayesian and Box-Jenkins procedures and compare their numerical results. Finally, the paper is concluded in Section VI.

II. AUTOREGRESSIVE PROCESSES

The autoregressive models are very useful in modeling time series data arise in many areas of scientific endeavor such as economics, business, marketing, physics, engineering and education.

Let $Y = [y(1)\ \ y(2)...y(n)]'$ be a vector of n observations generated from autoregressive process of order p, denoted by AR(p) for short. The model has the form (see Box and Jenkins (1970))

$$\phi(B)y(t) = \varepsilon(t) \qquad (1)$$

Where B is the backshift operator defined by

$$B^r y(t) = y(t-r) \quad , r = 1, 2, ...$$

y(t) denotes the time series observations, t=1, 2, …, n, ε(t) denotes the random errors assumed to be i. i. d. $N(0,\tau^{-1})$, $\tau^{-} > 0$ is the precision parameter. Moreover

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$$

The AR(p) model is always invertible and is stationary if the roots of the polynomial equation $\phi(B) = 0$ lie outside the unit circle. The model (1) can be written in more explicit form as

$$y(t) = \phi_1 y(t-1) + \phi_2 y(t-2) + ...$$
$$+ \phi_p y(t-p) + \varepsilon(t)$$

$$y(t) = \sum_{i=1}^{p} \phi_i y(t-i) + \varepsilon(t) \quad ,$$

$$t = ...,-2,-1,0,1,2,... \qquad (2)$$

The vector $\phi(p) = (\phi_1 \quad \phi_2 ... \phi_p)'$ is the vector of the unknown coefficients. In practice the order p is unknown and one has to estimate it using the vector of observations Y=[y(1) y(2) …y(n)]'. The Bayesian identification technique assumes that the order p is an additional parameter for which the marginal posterior probability mass function should be developed in a convenient form. The model (2.2) can be written in matrix notation as

$$y(p) = X(p)\phi(p) + \varepsilon(p) \qquad (3)$$

Where y(p) is a vector of order (n-p) with i-th element equal to y(p+i) and X(p) is a matrix of order (n-p)×p has the form

$X(p) = [y(t-1) \quad y(t-2) \ ...y(t-p)]'$ In addition, y(t-1) is a vector of order (n-p) with i-th element equal to y(p+i-1), y(t-2) is a vector of order (n-p) with i-th element equal to y(p+i-2), … and y(t-p) is a vector of order (n-p) with i-th element equal to y(i). The vector $\phi(p)$ is the vector of the unknown coefficients has the form $\phi(p) = [\phi_1(p) \quad \phi_2(p) \ .\phi_p(p)]'$ Finally, ε(p) is the vector the random errors of order (n-p) with i-th element equal to $\varepsilon(p+i)$. It is very important to mention that the vector y(p) and the matrix X(p) depend on the unknown order p. this means that for each value of p, say p_o, there is a corresponding vector $y(p_0)$ and a corresponding matrix $X(p_0)$.

III. BAYESIAN ANALYSIS OF NPO DATA

A. Identification

The time series of number of pilgrims coming from outside the kingdom of Saudi Arabia (NPO) (as shown in table (I))

consists of 44 observations (from year 1390AH[2] to year 1433 AH).

TABLE I. NUMBER OF PILGRIMS COMING FROM OUTSIDE THE KINGDOM (NPO)

Year	NPO	Year	NPO	NPO	Data	Year	NPO
1390	431270	1401	879368	1412	1012917	1423	1431012
1391	479339	1402	853555	1413	992813	1424	1419706
1392	645182	1403	1003911	1414	995611	1425	1534759
1393	607755	1404	919671	1415	1043274	1426	1557447
1394	918777	1405	851761	1416	1080465	1427	1654407
1395	894573	1406	856718	1417	1168591	1428	1707814
1396	719040	1407	960386	1418	1132344	1429	1729841
1397	739319	1408	762755	1419	1056730	1430	1613965
1398	830236	1409	774560	1420	1267555	1431	1799601
1399	862520	1410	828993	1421	1363992	1432	1828195
1400	812892	1411	720102	1422	1354184	1433	1752932

Let y(t) denotes the original series. We have found that the series y(t) is non-stationary and follows autoregressive scheme. The main objectives of this section is to identify the order of the series y(t), performing the diagnostic checking tests, and forecast the future observations using Bayesian approach. Given the assumptions outlined in the previous section, the conditional likelihood function of the process y(t) may be written as

$$L(\phi(p), p, \tau \mid y) \propto (\tau/2\pi)^{[n-p]/2} \exp(-\frac{\tau}{2}[y(p) - X(p)\phi(p)]'[y(p) - X(p)\phi(p)]),\ \phi(p) \in R^p,$$
$$\tau > 0,\ p = 1, 2, ..., k \qquad (4)$$

Where k is the known maximum value of the order of the process and X(p) is the same as defined in the previous section. Shaarawy and Ali (2003) developed the Bayesian identification analysis using a Normal. Gamma prior for the parameters $\phi(p)$ and τ . Here, we assume that the conditional prior distribution of the parameters $\phi(p)$ and τ given p is Jeffreys' non-informative prior defined by

$$g(\phi(p), \tau \mid p) \propto \tau^{-1},\ \tau > 0 \qquad (5)$$

With respect to the prior probability mass function of the order p, we will assume that

$$\beta_i = P_r[p = i] \quad , i = 1,2,..,k \qquad (6)$$

The joint posterior distribution of the parameters $\phi(p)$, p and τ is proportional to the multiplication of the conditional likelihood function (4) and the priors (5) and (6). Integrating the joint posterior distribution of the parameters with respect to $\phi(B)$ and τ , one may prove that the marginal posterior probability mass function of the order p is

$$h(p \mid y) \propto \beta_i(\pi)^{-v(p)} \left| X(p)X'(p) \right|^{-1/2}$$
$$\Gamma(v(p))\{y'(p)y(p) - y'(p)X(p)[X'(p)X(p)]^{-1}$$
$$X'(p)y(p)\}^{-r(p)/2} \qquad (7)$$

Where

$$2r(p) = n - 2p \quad , p = 1, 2, ..., 4$$

The formula (7) has been used with the following three priors for the order p (assuming k=4)

Prior 1: $\beta_i = 1/4 \quad , i = 1,2,..,4$

Prior2:

$$\beta_i \propto (0.5)^i \quad , i = 1,2,..,4$$

Prior3:

$$\beta_1 = 0.4\ , \beta_2 = 0.3,\ \beta_3 = 0.2\ ,\ \beta_4 = 0.1$$

The first prior assigns equal probabilities to the all possible values of the order p. The second prior is chosen in such a way to give probabilities that decline exponentially with the order, while the third prior is chosen in such a way to give probabilities that decrease with an amount 0.1 as the order increases. One may easily verify that the probabilities of the second prior are

$$\beta_1 = 0.5333\ ,\beta_2 = 0.2667$$
$$\beta_3 = 0.1333\ ,\beta_4 = 0.0667$$

The Matlab program has been used to do all computations required to calculate the marginal posterior probability mass function (7) for all three priors. The posterior probabilities are reported in table (II)

TABLE II. MARGINAL POSTERIOR PROBABILITY MASS FUNCTION OF THE ORDER FOR THE THREE PRIORS

Order	Jeffreys	Geometric	Arithmetic
1	1.0	1	1
2	0.0	0.0	0.0
3	0.0	0.0	0.0
4	0.0	0.0	0.0

[2] *The years ate written using Islamic (Lunar) Calendar (AH).For instance, the year 1433 corresponds to year2012A.D. For more details see http://en.wikipedia.org/wiki/Islamic_Calendar*

Table (II) shows that the marginal posterior function attains its maximum at the first order with a perfect probability for all three priors. This means that the tentative adequate model is AR(1) for the series y(t) regardless the used prior.

IV. DIAGNOSTIC CHECKING

The next phase of Bayesian time series analysis is to check the model, which has been tentatively identified as AR (1), to see if it gives a reasonable fit to the data at hand. This has been accomplished by doing three different types of tests. The first type contains the test concerning the significance of the coefficient ϕ_1. The second type is to do the over fitting test. The third type is to do the residual analysis using the estimated residuals. With regard to the first type, the absolute value of the estimated parameter was

TABLE III. THE VALUES OF THE χ^2 STATISTICS

Lag(k)	Box-Pierce statistic	Ljung-Box statistic
12	10.088	12.603
24	18.084	28.155
36	19.798	34.625

1.028. In addition, we have found that a 95% HPD interval for ϕ_1 was (0.995, 1.062). This means that we refuse the null hypothesis $H_0 : \phi_1 = 0$ and conclude that the parameter ϕ_1 is significant and the model AR (1) is appropriate for our data.

Regarding the over fitting test, the higher model AR (2) has been fitted to the data. We know that the marginal posterior distribution of the parameters ϕ_1 and ϕ_2 is a non-central t with (n-4) degrees of freedom and location and precision parameters given by Broemeling and Shaarawy (1988). We have found that a 95% HPD interval for the added parameter ϕ_2 is (-0.027, 0.627) which conclude the zero value. Thus we cannot refuse the null hypothesis $H_0 : \phi_2 = 0$ and conclude that the identified model AR (2) is not appropriate for the data.

The third type diagnostic checking is to do residual analysis. If the fitted model AR(1) is appropriate, the calculated residuals $\hat{\varepsilon}(1), \hat{\varepsilon}(2), ..., \hat{\varepsilon}(n)$ should behave in manner which is consistent with the true model. This has been accomplished by doing several checks such as time series plot, the autocorrelation and partial autocorrelation functions of the residual, the portmanteau lack of fit test, and the autocorrelation function of the first difference of the residual. However, the time series plot of the residuals shows no outliers or any non-desirable autocorrelation or cyclic effect. The plot also gives no indication of a non-zero mean or non-constant variance. In addition, the autocorrelation and partial autocorrelation function of the residuals have no spikes. Moreover, The Anderson-Darling statistic for testing the normality assumption is 0.408 with p-value 0.333.These mean that the residuals resemble that of a whit noise sequence which supports the appropriateness of the identified model AR(1).

Instead of testing each autocorrelation, it is recommended to inspect the first k autocorrelation of the residual simultaneously using the Box-Pierce or Ljung-Box statistics. These two statistics have been calculated and the results are reported in table (III).

Comparing the results given by table (III) with the critical values of the χ^2 distribution with (k-1) degrees of freedom, we do not reject the null hypothesis

$$H_0 : \rho_\varepsilon(1) = \rho_\varepsilon(2) = ... = \rho_\varepsilon(k) = 0$$

for all values of k. These results support the appropriateness of the identified model AR(1). For more details about those two statistics, the reader is referred to Box and Jenkins (1970). Finally, the graph of the autocorrelation function of the first difference of the residuals cuts off after the first lag, while its partial autocorrelation function decays down. This means that the series of the first difference of the residuals has an MA(1) model with parameter does not differ significantly from1(see Box and Jenkins(1970)). This gives another support to the identified model AR(1).

B. Forecasting

The last phase of time series analysis is to forecast future observations. Thus, after passing through the modeling and diagnostic checking tests, confident that an AR(1) process has generated the NPO series, one would like to forecast Y(n+1), Y(n+2), …. The posterior predictive density of the future observations is the Bayesian tool to solve the forecasting problems.

a) One Step-Ahead Predictive Distribution

Assuming Jeffreys' non-informative prior (3.2) for the parameters ϕ_1 and τ , the predictive density of the next future observation y(n+1) can be shown to follow a non-central t distribution with (n-2) degrees of freedom, location

$$E[W(n+1) \mid y] = DE,$$

And precision

$$P[W(n+1) \mid Y] = D(n-2)(F - \frac{E}{D})^{-1}$$

Where

$$D = 1 - \frac{w^2(n)}{\sum_{t=1}^{n} w^2(t)},$$

$$E = \frac{w(n)\sum_{t=2}^{n} w(t)w(t-1)}{\sum_{t=1}^{n} w^2(t)} \quad \text{and}$$

$$F = \sum_{t=2}^{n} w^2(t) - \frac{[\sum_{t=2}^{n} w(t)w(t-1)]^2}{\sum_{t=1}^{n} w^2(t)}$$

The posterior mean $E[Y(n+1) \mid y]$ provides a point forecast for the next observation y(n+1), and a $(1-\alpha)\%$ HPD interval for y(n+1) is

$$E[Y(n+1) \mid y) \pm t_{\alpha/2,n-2} P^{-1/2}[Y(n+1) \mid y]$$

b) Multi-Step-Ahead Predictive Distribution

The procedure followed throughout the above subsection to predict the first future observation y(n+1), using the one step-ahead predictive density, can be generalized. So that we can predict the k^{th} future observation y(n+k), using the k step-ahead predictive density. However, the prediction process of y(n+k) is conditional on the predictions of its preceding future observations y(n+1), y(n+2),.., y(n+k-1).

Thus, the forecasting process of the future observations should be employed step by step. One should first predict y(n+1) using the one step-ahead predictive density. Then, depending on a point forecast for y_{n+1} one can predict y(n+2) using the two step-ahead predictive density, which is conditional on the point forecast of y(n+1).

This process can be repeated for the succeeding future observations. For more details, see Broemeling and Shaarawy (1988).The model AR (1) has been used to forecast the next five future observations. The point forecasts and 95% HPD intervals for these observations are given by table (IV).

TABLE IV. THE FUTURE FIVE FORECASTS AND THEIR CONFIDENCE INTERVALS USING BAYESIAN PROCEDURE.

Year	Point forecast	HPD intervals
1434	1790544	(1571767, 2009320)
1435	1828962	(1612811, 2045113)
1436	1868204	(1654586, 2081823)
1437	1908289	(1697115, 2119464)
1438	1949234	(1740421, 2158047)

V. BOX AND JENKINS ANALYSIS OF NPO DATA

Box and Jenkins (1970) have presented a statistical analysis of ARMA(p,q) processes which has grown in popularity and is today the prevailing methodology of time series analysis. They assume that the time series at hand (or a transformation of the series) could be presented by a parsimonious stationary and invertible ARMA process such that one can perform the four phases of time series analysis: identification (order determination), estimation, diagnostic checking, and forecasting. In what follows we give a brief summary to each phase using Box and Jenkins methodology.

According to Box and Jenkins, the identification of the order p and q is done by computing the sample autocorrelation and partial autocorrelation functions and matching them with their theoretical counterparts, which are mathematically known for low-order processes.

Their methodology has been widely used and explained by many others such as Chatfield (1980), Priestely (1981), Bowerman and O'Connell (1987), Tong (1990), Harvey (1993), Wei (2005), Box et al.(2008) and Liu (2009).

After the model is tentatively identified, say an ARMA (p, q), the autoregressive parameters $\phi = (\phi_1 \quad \phi_2 \quad \dots \quad \phi_p)^T$, the moving average parameters $\theta = (\theta_1 \quad \theta_2 \quad \dots \quad \theta_q)^T$, and the residual variance σ^2 are estimated by maximum likelihood or nonlinear least squares methods. The maximum likelihood estimates (MLE) and least squares estimates (LSE) may be based on either the full (unconditional) likelihood function or a conditional likelihood function. These techniques are given in details by Priestely (1981). If q=0, the noise term is a linear function of the parameters and one may use the well-known linear least squares algorithm to estimate the parameters $\phi_1, \quad \phi_2, \dots \quad \phi_p$.

The third phase of a time series is to check the adequacy of the identified model to see if it gives a reasonable fit to the data at hand. This is mainly accomplished by a series of diagnostic checks using the estimated residuals. One may inspect the graphs of autocorrelation and partial autocorrelation to make sure that they do not have significant spikes particularly at low lags. In addition, one may investigate the residual plot to make sure that it does not have a particular pattern. Moreover, one may investigate the Ljung and Pierce statistic. One may also investigate the fitted model of the first differences of the residuals to see if it has the first order moving average model. For more details, see Box and Jenkins (1976) and Box et al. (2008).

The last phase of a time series analysis is to forecast future observations where the predicted observations are computed recursively from an estimated conditional expectation, namely, the conditional expectation of a future observation given the past data. For more details see, Box and Jenkins (1976) and Box et al. (2008).

The main objective of first section is to model and forecast the NPO data using the most popular well - known approach developed by Box and Jenkins in 1970. In order to achieve an adequate tentative model for the NPO data, the time series plot and the autocorrelation function (acf) are plotted in figures (I) and (II) respectively.

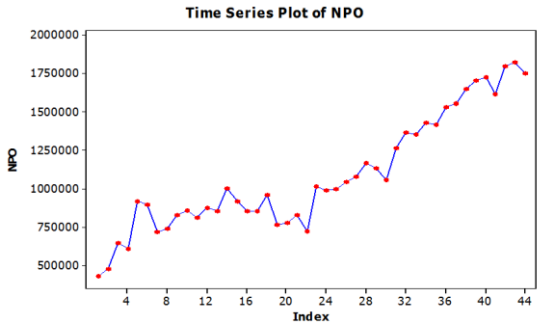


Fig. 1.

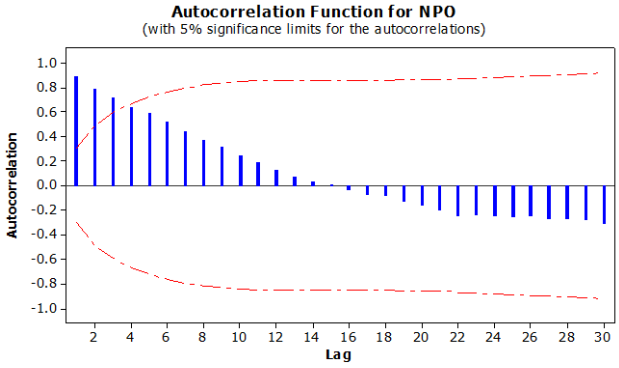


Fig. 2.

Inspecting the above graphs, it is easy to conclude that the NPO data y_t is non - stationary in the mean and variance.

In order to use Box and Jenkins methodology, it was necessary to use some mathematical transformation to convert the original data y_t into a new stationary series. As we have said before, this one of the disadvantages of Box and Jenkins methodology. After doing many trails, we can say that the second difference of the logarithm of the NPO data succeeded to convert the original time series y_t to a stationary one. Let z_t denotes to the new series, then

$$z_t = \log(y_t) - 2\log(y_{t-1}) + \log(y_{t-2}), \quad t = 3,4,...,44$$ (III)

Figures and (IV) show the time series plot and the autocorrelation function of the time series z_t .

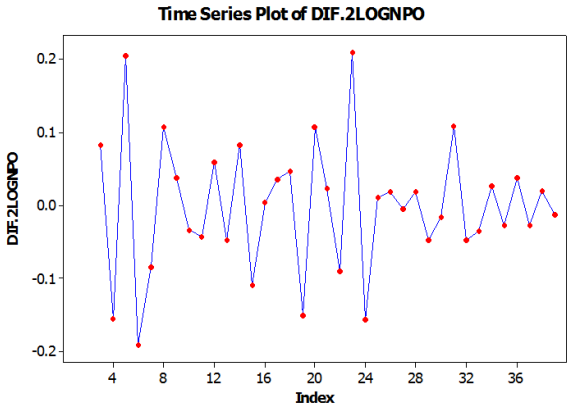


Fig. 3.

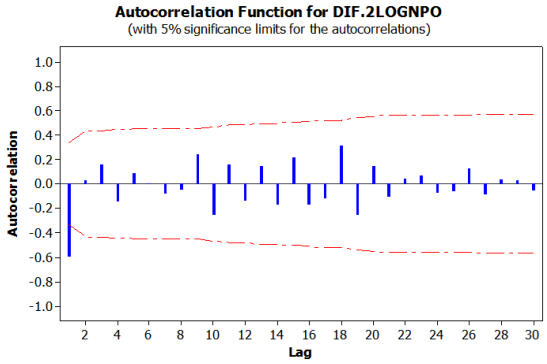


Fig. 4.

Inspecting these graphs, one may say that the time series z_t tends to be stationary. Thus, one may use Box and Jenkins to model and forecast the series z_t . In order to identify a tentative model to z_t data, the partial autocorrelation function is computed and its graph is given by figure (V).

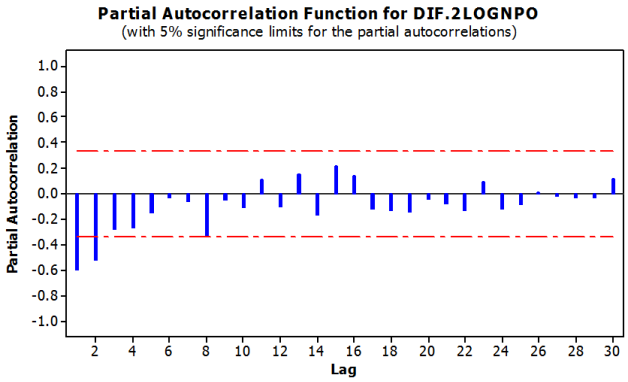


Fig. 5.

Inspecting the autocorrelation of the series z_t , one may notice that the coefficients of the autocorrelation function are small after the first lag. In addition, one may notice that the partial autocorrelation coefficients are small at the third and fourth lags and very small after the fourth lag. This means that we have four different models, one of them must be chosen in order to have good forecasts. The first choice is the first order moving average, denoted by MA(1), model , while the other three choices are AR(2), AR (3), and AR (4) models. We have started to analyze the series z_t using MA(1) model, but the numerical results were unsatisfactory because the model did not pass most of the diagnostic checking tests.

For example the autocorrelation function seems to have a spike at the first lag and the p-value for Anderson-Darling statistic was 0.014. Second, the AR(2) model was used to fit the data, but the numerical results of the diagnostic checking have shown that more autoregressive parameters should be added to the model. Therefore, we have used AR(3) model to fit the data, but it turned out that the model still needs more autoregressive parameters. Finally AR (4) model has been used to analyze the data, and all numerical results of the diagnostic tests were very satisfactory. Thus, we concluded that AR (4) model is the most adequate one the model and forecast the series Z_t , i.e. the ARIMA (4, 2, 0) is the most adequate model to fit the logarithm of the NPO data.

Finally, the model ARIMA(4,2,0) has been used to forecast the next five future observations. The point forecasts and 95% HPD intervals for these observations are given by table (V).

TABLE V. THE FUTURE FIVE FORECASTS AND THEIR CONFIDENCE INTERVALS USING BOX AND JENKINS PROCEDURE

Year	Point forecast	Confidence intervals
1434	1827212	(1423122 , 2346042)
1435	1828925	(1316379 , 2541038)
1436	1880008	(1247717 , 2832718)
1437	1896373	(1109448 , 3241462)
1438	1893155	(979379 , 3659497)

VI. A COMPARATIVE STUDY

This section has three main objectives. The first is to evaluate the forecast performance of Bayesian procedure outlined in section 3. The second objective is to evaluate the forecast performance of the Box-Jenkins procedure used in section 4.

The final objective is to compare the numerical results achieved by the two proposed approaches. In order to achieve the main goals, a small portion of the NPO data at the end of the data are reserved solely for forecast comparison. In statistical literature, these data are referred to as *hold-out sample, or post-sample,* and in principle are not used in model identification or estimation when evaluating forecast performance. However, there are several criteria to evaluate forecast performance of a model, including mean absolute percentage error (MAPE), mean absolute deviation (MAD), and root mean squared error (RMSE) as defined below

$$MAPE = \frac{1}{m}\sum_{t=1}^{m}\left|\frac{real\ value - forecast}{real\ value}\right|.100$$

$$MAD = \frac{1}{m}\sum_{t=1}^{n}\left|true\ value - forecast\right|$$

$$RMSE = \left[\frac{1}{m}\sum_{t=1}^{n}\left(true\ value - forecast\right)^2\right]^{1/2}$$

Where, m is the total number of observations in the *hold-out sample (post-sample).*

In order to use the above criteria to evaluate the forecast performance of the two proposed approaches and compare between them, the last 5 observations (about 12% of the whole data) are reserved as the *hold-out sample (post-sample).* The first 39 observations were used to forecast the next five observations using Bayesian and Box-Jenkins approaches; then the five forecasts were compared with the five real observations and the three above criteria were calculated. The results are reported in tables (VI) and (VII) respectively.

TABLE VI. THE FORECAST PERFORMANCE OF BAYESIAN APPROACH

Year	Real observations	Forecasts	APE%	AD	SE
1429	1707814	1756113	1.52	48299	2332793401
1430	1729841	1805778	11.88	75937	5766427969
1431	1613965	1856848	3.18	242883	58992151689
1432	1799601	1909363	4.44	109762	12047696644
1433	1828195	1963362	12.00	135167	18270117889
Mean			5.7000	122409.6	19481837518.4

TABLE VII. THE FORECAST PERFORMANCE OF BOX-JENKINS APPROACH

Year	Real observations	Forecasts	APE%	AD	SE
1429	1707814	1748546	1.08	40732	1659095824
1430	1729841	1852292	14.77	122451	14994247401
1431	1613965	1904772	5.84.	290807	84568711249
1432	1799601	1985503	8.61	185902	34559553604
1433	1828195	2066444	17.89	238249	56762586001
Mean			9.600	175628.2	38508838815.8

Inspecting the results given by table (I), one may conclude that the identified model AR(1) gives very good Bayesian forecasts since the mean absolute percentage error (MAPE) is very low, being equal to 5.7%. On the other hand, the corresponding value of MAPE computed by Box-Jenkins procedure is 9.6%. In addition the MAD and the RMSE for Bayesian approach were 122409.6 and 139577.4 respectively, while the corresponding values for Box-Jenkins approach were 175628.2 and 196236.7.

In addition, 95% confidence intervals intervals have been computed for the *hold-out sample (post-sample)* using the two proposed procedures and the results are reported in table (VIII) and (IX).

TABLE VIII. 95% CONFIDENCE INTERVALS OF THE LAST FIVE OBSERVATIONS USING BAYESIAN ANALYSIS

Year	Lower Bound	Upper Bound	Length of the Intervals
1429	1655613	2094008	438395
1430	1706121	2139107	432986
1431	1757749	2185525	427776
1432	1810534	2233287	422753
1433	1864513	2282418	417906
Mean			427962

TABLE IX. CONFIDENCE INTERVALS OF THE LAST FIVE FUTURE OBSERVATIONS USING BOX AND JENKINS PROCEDURE

Year	Lower Bound	Upper Bound	Length of the Intervals
1429	1423122	2346042	922920
1430	1316379	2541038	1224659
1431	1247717	2832718	1585001
1432	1109448	3241462	2132014
1433	97379	3659497	2680119
Mean			1708943

Comparing the numerical results of Bayesian approach, in forecasting the last five observations, with the numerical results of Box and Jenkins approach, one may conclude the flowing:

1) The mean absolute percentage error (MAPE) achieved by Box and Jenkins approach is higher than the corresponding value achieved by Bayesian approach with more than 68%.

2) The mean absolute deviation (MAD) achieved by Box and Jenkins approach is higher than the corresponding value achieved by Bayesian approach with more than 43%.

3) The root mean squared error (RMSE) achieved by Box and Jenkins approach is higher than the corresponding value achieved by Bayesian approach with more than 40%.

4) The 95% confidence interval for the next step ahead forecast achieved by Box and Jenkins approach is wider than the corresponding value achieved by Bayesian approach with more than 110%.

5) The mean of lengths of the 95% confidence intervals for the last five observations achieved by Box and Jenkins approach is higher than the corresponding value achieved by Bayesian approach with more than 299%.

From the foregoing numerical results, we conclude that the Bayesian approach is much more accurate than Box and Jenkins approach in modeling and forecasting the NPO data because it gives better forecasts and narrower confidence intervals.

VII. SUMMERY AND CONCLUSION

The authors have proposed to use the Bayesian approach to develop a complete time series analysis of number of Pilgrims coming from outside the Kingdom of Saudi Arabia from year 1390AH to year 1433AH. Using a Jeffreys' non-informative prior for the parameters and three different priors for the model order, the proposed methodology is to develop the marginal posterior probability mass function of the model order is given in an easy and convenient form using the approach developed by Shaarawy and Ali(2003) . Then, one may investigate the behavior of the marginal posterior probability mass function and choose the order at which the marginal probability mass function attains its maximum to be the identified order. We have found that AR(1) model is the identified tentative model for the series. The tentative model has passed all the diagnostic checking tests with high precision. Point forecasts and HPD intervals for the next five future years are provided by the authors using the marginal and conditional predictive densities given by Broemeling and Shaarawy (1988). In addition, the traditional Box and Jenkins approach was used to analyze the same data. The numerical results achieved by Bayesian approach were much better than the results achieved by the traditional Box and Jenkins approach.

REFERENCES

[1] Bowerman, B. and O'Connell (1987). Time Series Forecasting: Unified Concepts and Computer Implementation. Boston: PWS Publishers Dexbury Press.

[2] Box, G. and Jenkins, G. (1970). Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco

[3] Broemling, L. and Shaarawy, S. (1988). Time Series Analysis: A Bayesian Analysis in the Time Domain. Bayesian Analysis of Time Series and Dynamic Model, edited by Spall, J.

[4] Chatfield, C. (1980). The Analysis of Time Series: Theory and Practice. Chapman and Hill Ltd, London.

[5] Chatfield, C. (2004). The Analysis of Time Series: An Introduction. 6th ed. Boca Raton, Fla. : Chapman & Hall/CRC, - Texts in statistical science.

[6] Chow, G. (1975). Multi-period Prediction from Stochastic Difference Equations by Bayesian Methods. Chapter 8 of Studies in Bayesian Econometrics and Statistics, edited by S.E. Fienberg and A. Zellner. North-Holland, Amsterdam.

[7] Harvey, A. (1993). Time Series Models, 2nd edition. The MIT Press.

[8] Liu, L. (2009), Time Series Analysis and Forecasting. 2nd edition. Scientific Computing Association Corp, USA.

[9] Monahan, J. (1983). Fully Bayesian Analysis of ARIMA Time Series Models, Journal of Econometrics, Vol. 21, pp. 307-331.

[10] Newbold, P.(1973). Bayesian Estimation of Box and Jenkins Transfer Function Model for Noise Models. Journal of the Royal Statistical Society, Series B, vol. 35. No. 2, pp. 323-336.

[11] Priestley, M. (1981). Spectral Analysis of Time Series, Academic Press, London.

[12] Shaarawy, S. and Ali, S. (2003). Bayesian Identification of Seasonal Autoregressive Models. Communications in Statistics-Theory and Methods, Vol. 32, Issue 5, pp.1067-1084.

[13] Shaarawy, S. and Ali, S. (2008). Bayesian Identification of Multivariate Autoregressive Processes. Communications in Statistics-Theory and Methods, Vol. 37, Issue 5, pp.791-802.

[14] Shaarawy, S. and Broemeling, L. (1984). Bayesian Inference and Forecasts with Moving Average Processes. Comm. In Statis. Vol.13, No. 15.

[15] Shaarawy, S., Soliman, E. and Ali, S. (2007). Bayesian Identification of Moving Average Models, Communications in Statistics-Theory and Methods, Vol. 36, Issue 12, pp. 2301-2312.

[16] Tong, H. (1990). Non-Linear Time Series: A Dynamical System Approach. New York: Oxford University Press.

[17] Wei, W.W.S. (2005). Time Series Analysis: Univariate and Multivariate Methods. Addison Wesley, Reading, MA.

[18] Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. John Wiley and Sons. Inc, New York.

[19] Zellner, A. and Reynolds, R. (1978). Bayesian Analysis of ARMA Models. Presented at the Sixteenth Seminar on Bayesian Inference in Econometrics, June 2-3.

A new Hierarchical Group Key Management based on Clustering Scheme for Mobile Ad Hoc Networks

Ayman EL-SAYED, IEEE Senior Member
Department of Computer Science and Engineering,
Faculty of Electronic Engineering,
Menoufiya University, Menouf 32952, Egypt.
Email: ayman.elsayed@el-eng.menofia.edu.eg

***Abstract*—The migration from wired network to wireless network has been a global trend in the past few decades because they provide anytime-anywhere networking services. The wireless networks are rapidly deployed in the future, secure wireless environment will be mandatory. As well, The mobility and scalability brought by wireless network made it possible in many applications. Among all the contemporary wireless networks,Mobile Ad hoc Networks (MANET) is one of the most important and unique applications. MANET is a collection of autonomous nodes or terminals which communicate with each other by forming a multihop radio network and maintaining connectivity in a decentralized manner. Due to the nature of unreliable wireless medium data transfer is a major problem in MANET and it lacks security and reliability of data. The most suitable solution to provide the expected level of security to these services is the provision of a key management protocol. A Key management is vital part of security. This issue is even bigger in wireless network compared to wired network. The distribution of keys in an authenticated manner is a difficult task in MANET. When a member leaves or joins the group, it needs to generate a new key to maintain forward and backward secrecy. In this paper, we propose a new group key management schemes namely a Hierarchical, Simple, Efficient and Scalable Group Key (HSESGK) based on clustering management scheme for MANETs and different other schemes are classified. Group members deduce the group key in a distributed manner.**

Keywords– Group Key management; Mobile Ad hoc network; MANET security; Unicast/Multicast protocols in MANET.

I. Introduction

Mobile Ad Hoc Network (MANET) [1], [2] is kind of mobile, multiple hops, and self-discipline system, not depend on the fixed communication facilities. Ad Hoc network is a series of nodes in structure which move anywhere at will, the network nodes distribute dynamically, nodes contact others through wireless network, every network node has the double functions as terminal and routers, the nodes are peer-to-peer, communicate with a high degree of coordination. Wireless Ad Hoc network is flexibility with a wide foreground of application, mainly used in multimedia conference, emergency rescue, relief, exploration, military action and sensor network etc. [3]. A communication session is achieved either through single-hop transmission if the recipient is within the transmission range of the source node, or by relaying through intermediate nodes otherwise. For this reason, MANETs are also called multi-hop packet radio network [4], [5]. However, the transmission range of each low-power node is limited to each other's proximity, and out-of-range nodes are routed through intermediate nodes. On the contrary to traditional network architecture, MANET does not require a fixed network infrastructure; every single node works as both a transmitter and a receiver. Nodes communicate directly with each other when they are both within the same communication range. Otherwise, they rely on their neighbors to relay messages. The self-configuring ability of nodes in MANET made it popular among critical mission applications like military use or emergency recovery. However, group key management for large and dynamic groups in MANETs is difficult problem because of the requirement of scalability, security under the restrictions of nodes' available resources and unpredictable mobility [6]. But the group key management protocols dedicated to operate in wired networks are not suited to MANET, because of the characteristics and the challenges of such environments [7]. So many researchers are interesting of group key management for MANET. In our issue, group key management means that multiple parties need to create a common secret to be used to exchange information securely. Without central trusted entity, two people that have not previously a common share key can create a key based on the Diffie-Hellman (DH) protocol [8]. DH key agreement requires that both the sender and recipient of a message have key pairs. By combining one's private key and the other party's public key, both parties can compute the same shared secret number. This number can then be converted into cryptographic keying material. It is called 2-party DH protocol that can be extended to a generalized version of n-party DH. In [9], the authors integrated the DH key exchange into the Digital Signature Algorithm (DSA) and in [10], the authors fix this integration protocols so that both forward secrecy and key freshness can be guaranteed, while preserving the basic essence of the original protocols. This fix also provides key freshness because every session key is a function of ephemeral secrets chosen by both parties, so neither party can predetermine a session key's value since he would not know what the other party's ephemeral secret is going to be. However, robust key management services are central to ensuring privacy protection in wireless ad hoc network settings. Existing approaches to key management, which often rely on trusted, centralized entities, are not well–suited for the highly dynamic, spontaneous nature of ad hoc networks. So many researchers are interesting to make proposals for key management techniques that are

surveyed in [11] to find an efficient key management for secure and reliable. This paper proposes one of the group key management schemes namely a Hierarchical, Simple, Efficient and Scalable Group Key (HSESGK) based on clustering management scheme for MANETs. Group members compute the group key in a distributed manner. This hierarchical contains two levels only, first level for all coordinators of the clusters as a main group's members; it is called cluster head (CH) that is selected by the algorithms shown in [12], [13], [14], the second level for the members in a cluster with its cluster head. Then there are two secret keys obtained in a distributed manner, the first key among all the CHs and the second key among cluster's members and its CH. HSESGK uses double trees in each cluster for robustness and avoid fault tolerance. Also group key management is to ensure scalable and efficient key delivery, taking into account the node mobility.

The remainder of this paper is organized as follows: Section II reviews related work such that MANET routing protocols for both unicast and multicast and security requirements. Also this section describes the overview of MANET key management and short note about our proposal. Details of our group key management scheme are described in Section III and our scheme is discussed with some features in Section IV. Finally, we conclude the paper in Section V.

II. Related work

A. MANET unicast routing protocols

Several routing protocols [15] have been proposed in recent years for possible deployment of Mobile Ad hoc Networks (MANETs) in military, government and commercial applications. In [16], these protocols are reviewed with a particular focus on security aspects. The protocols differ in terms of routing methodologies and the information used to make routing decisions. Four representative routing protocols are chosen for analysis and evaluation including: Ad Hoc on demand Distance Vector routing (AODV), Dynamic Source Routing (DSR), Optimized Link State Routing (OLSR) and Temporally Ordered Routing Algorithm (TORA). Secure ad hoc networks have to meet five security requirements: confidentiality, integrity, authentication, non-repudiation and availability. The analyses of the secure versions of the proposed protocols are discussed with respect to the above security requirements. Routing protocols for ad hoc wireless networks can be classified into three types based on the underlying routing information update mechanism employed as shown in Fig. 1. An ad hoc routing protocol could be reactive (on demand), proactive (table driven) or hybrid.

Reactive routing protocols obtain the necessary path, when required, by using a connection establishment process. Such protocols do not maintain the network topology information and they do not exchange routing information periodically. In this section, we will focus on three routing protocols and some of their secure versions. First, we discuss DSR [17]. The secure versions, such as, QoS Guided Route Discovery [18], Securing Quality of Service Route Discovery [19], Ariadne [20] and CONFIDANT [21] are presented as well. Second, AODV [22] is discussed with its secure versions, CORE [23],

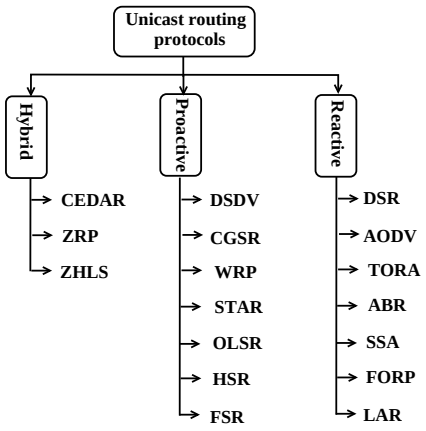


Fig. 1. Ad hoc unicast routing protocols

SAODV [24] and SAR [25]. Finally, TORA [26] is discussed followed by the discussion of two ad hoc security techniques, SPREAD [27] and ARAN [28]. We focus more on reactive routing protocols because they often outperform proactive ones due to their ability to adjust the amount of network overhead created to track the mobility in the network affecting current communication.

In ***proactive or table driven routing protocols***, such as DSDV [29] or OLSR [30], every node maintains the network topology information in the form of routing tables by periodically exchanging routing information. Routing information is generally flooded in the whole network. Whenever a node requires a path to a destination, it runs an appropriate path finding algorithm on the topology information it maintains.

Hybrid routing protocols such as ZRP [31] and SRP [32] are protocols that combine the best features for both reactive and proactive routing protocols. For example, nodes communicate with their neighbors using proactive routing protocols and communicate with far distance nodes using reactive routing protocols.

B. MANET Multicast routing protocols

There is a need for multicast traffic also in ad hoc networks. The value of multicast features with routing protocols is even more relevant in ad hoc networks, because of limited bandwidth in radio channels [33]. Some multicast protocols [34], [35] are based to form and maintain a routing tree among group of nodes. Some other are based on to use routing meshes that have more connectivity than trees etc.

The various classifications of the multicast routing protocols in MANETs are shown in Fig. 2. It illustrates the main classification dimensions for multicast routing protocols such as: *multicast topology*, *initialization approach*, *routing scheme*, and *maintenance approach*.

Multicast topology [36]: it is classified into two approaches namely mesh based and tree based [37], [38]. Tree based approach is classified into two types: *Source tree based*, in which each source creates a separated tree that contains the source as a root of the tree. *Shared tree based*, in which

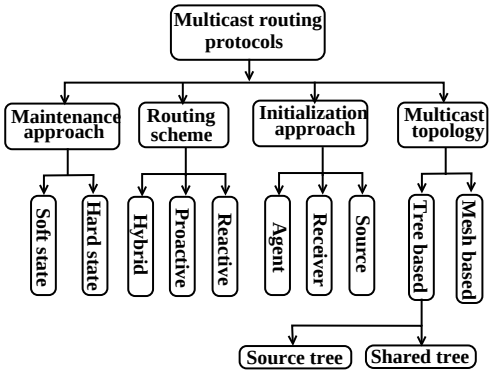


Fig. 2. Ad hoc multicast routing protocols

one tree is created in the entire network which includes all sources and receivers. Mesh based approach depends on multiple paths between any source and receivers pair. The mesh based protocols create the tree dependent on the mesh topology. These redundant paths are useful in link failure case and provide higher packet delivery ratio.

Routing initialization approach: Routing initialization is classified into three approaches namely source-initiated, receiver-initiated, and hybrid approach [39]. *Source initiated*: the source is responsible of construction and maintenance the group tasks. *Receiver initiated*: the receiver searches the multicast group to join with dedicated source. *Hybrid initiated*: the multicast group construction and maintenance tasks are done by either the source or the receiver.

Routing scheme: Routing scheme is classified into three approaches namely table-driven (proactive), on-demand (reactive), and hybrid approach [38], [39] as the same meaning in the unicast routing protocols explained in previous section.

Maintenance approach: Multicast maintenance is classified into two approaches namely soft-state and hard-state. *Soft-state approach*: a route maintenance process initiated periodically by flooding the network with control packets to explore other routes between source and receiver. This approach has the advantage of reliability and better packet delivery ratio, but it is much makes overhead over the network as it continuously floods the network with control packets [39]. *Hard-state approach*: a route maintenance process is established by two types namely reactive and proactive. In reactive approach, broken link recovery process is initiated only when a link breaks. The second type is proactive approach, in which routes are reconfigured before a link breaks, and this can be achieved by using local prediction techniques based on GPS or signal strength [39].

C. Security Requirements

The security services of ad hoc networks are not different of those of other network communication paradigms. Specifically, an effective security paradigm must ensure the following security primitives: *identity verification*, *data confidentiality*, *data integrity*, *availability*, and *access control*. Although solutions to the above concerns have been developed and widely deployed in the wired domain, the amorphous, transient properties of ad hoc networks preclude their adaptation to serverless network environments, which are often comprised of small devices. Instead, security solutions, in general, and key managements should strive for the following characteristics:

***Lightweight*:** Solutions must minimize the computation and communication processing required to ensure the security services to accommodate the limited energy and computational resources of ad hoc enabled devices.

Decentralized: Like ad hoc networks themselves, attempts to secure them must be ad hoc: they must establish security without a priori knowledge or reference to centralized, persistent entities. Instead, security paradigms must levy the cooperation of all trustworthy nodes in the network.

***Reactive*:** Ad hoc networks are dynamic: nodes trustworthy and malicious may enter and leave the network spontaneously and unannounced. Security paradigms must react to changes in network state; they must seek to detect compromises and vulnerabilities; they must be reactive, not protective.

***Fault–Tolerant*:** Wireless transfer mediums are known to be unreliable; nodes are likely to leave or be compromised without warning. The communication requirements of security solutions should be designed with such faults in mind; they mustn't rely on message delivery or ordering.

D. MANET key management overview

MANET has some constrains such its energy constrained operations, limited physical security, variable capacity links and dynamic topology. So, there are different Key Management schemes are used to achieve the high security in using and managing keys. The crucial task in MANET uses different cryptographic keys for encryption like symmetric key, asymmetric key, group key and hybrid key (i.e. mixed of both symmetric key and asymmetric key). Here we discuss about some of the important Key Management schemes in MANET and they are shown in Fig. 3.

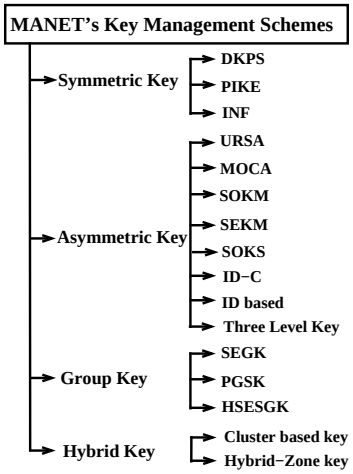


Fig. 3. Key Management Schemes in MANET

1) Symmetric Key Management: In symmetric key management, the same keys are used by sender and receiver. This key is used for encryption the data as well as for decryption the data. If n nodes wants to communicate in MANET, k number of key pairs are required, where k=n(n-1)/2. Some of the symmetric key management schemes in MANET are Distributed Key–Pre Distribution Scheme (DKPS) [40], Peer Intermediaries for Key Establishment (PIKE) [41], and Key Infection (INF) [42].

2) Asymmetric Key Management: Asymmetric keys uses two-part key. Each recipient has a private key that is kept secret and a public key that is published for everyone. The sender looks up or is sent the recipient's public key and uses it to encrypt the message. The recipient uses the private key to decrypt the message and never publishes or transmits the private key to anyone. Thus, the private key is never in transit and remains invulnerable. This system is sometimes referred to as using public keys. This reduces the risk of data loss and increases compliance management when the private keys are properly managed. Some of the asymmetric key management schemes in MANET are Self-Organized Key Management (SOKM) [43], Secure and Efficient Key Management (SEKM) [44], Private ID based Key Asymmetric Key Management Scheme [45].

3) Group Key Management Scheme: Group key in cryptography is a single key which is assigned only for one group of mobile nodes in MANET. For establishing a group key, group key is creating and distributing a secret for group members. There are specifically three categories of group key protocol. (1) Centralized, in which the controlling and rekeying of group is being done by one entity. (2) Distributed, group members or a mobile node which comes in group are equally responsible for making the group key, distribute the group key and also for rekeying the group. (3) Decentralized, more than one entity is responsible for making, distributing and rekeying the group key. Some important Group key Management schemes in MANET are Simple and Efficient Group Key Management (SEGK) [46], and Private Group Signature Key (PGSK) [47].

4) Hybrid Key Management Schemes: Hybrid or composite keys are those key which are made from the combination of two or more than two keys and it may be symmetric or a asymmetric or the combination of symmetric & asymmetric key. Some of the important Hybrid key management schemes in MANET are Cluster Based Composite Key Management [48], [49], and Zone-Based Key Management Scheme [50].

5) Our approach: In this paper, we propose the network model that contains some clusters; each cluster has its coordinator namely cluster head (Cluster initiator). The clusters are interconnected via the cluster heads. There are subgroups of members called cluster in which one member is cluster head and virtual subgroup of clusters' heads. Our model seems like Cluster-Head Gateway Switch Routing (CGSR) Protocol [51], [52] but in multicast manner, an optimized cluster based approach for multi-source multicast routing protocol in MANET [53] and Cluster Based Routing Protocol (CBRP) [54]. Our new key management scheme namely "Hierarchical, Simple, Efficient and Scalable Group Key based on clustering" (HSESGK) scheme that has main idea shown in [55]. The basic idea of our scheme is that a multicast tree is formed in MANETs for efficiency. A multiple tree based multicast routing scheme are used as mentioned in [56], [57], which exploit path diversity for robustness. Also in [46], the author used two multicast trees for improving the efficiency and maintains it in parallel fashion to achieve the fault tolerances. So, in our scheme, two multicast trees are used for each subgroup (i.e. cluster subgroups or cluster heads' subgroup). For example, in a cluster, the connection of multicast tree is maintained be its cluster head that compute and distribute the intermediate keying materials to all members in this cluster through the active tree links. Also the cluster head is responsible for maintaining the connection of the multicast subgroup. In MANET, main cluster head (MANET initiator) has the same role of cluster head, but on the clusters' subgroup.

III. Our Group key management scheme

A. Notations and assumptions

Firstly, every node takes a valid certificate from offline configuration before entering the network. An underlying public key infrastructure is then used to manage certificates. However, many researchers are interesting of this hot topic, and most key management proposals suffer the man-in-the-middle attack. In this paper, each member has a unique identifier and all keying materials signed by the coordinator (i.e. cluster head) in subgroup to make sure authenticity and integrity, in order to avoid the man-in-the-middle attack. Also, a group member has a password to join or can present a valid certificate. In our work, a group member can join by using a valid certificate. Here, for simplicity, we assume that a node can join a group if it has a valid certificate. Some notations used in HSESGK are listed as follows:

$\boldsymbol{M}_i$	i^{th} group member.
$\boldsymbol{g}$	Exponentiation base.
$\boldsymbol{p}$	Prime value.
$\boldsymbol{CH}_i$	i^{th} Cluster Head.
$\boldsymbol{MCH}$	Main Cluster Head.
$\boldsymbol{N}$	Total number of group members.
$\boldsymbol{N}_c$	Total number of Clusters.
$\boldsymbol{n}_{ci}$	Number of group members in i^{th} Cluster.
$\boldsymbol{r}_i$	A random number generated by i^{th} member, also called member private key.
$\boldsymbol{br}_i$	Blinded i^{th} member key. $\boldsymbol{br}_i = \boldsymbol{(g)}^{ri} \boldsymbol{\ mod\ p}$
$\boldsymbol{k}_i$	Internal i^{th} member key, or intermediate key.$\boldsymbol{k}_i = \boldsymbol{(b}_{ri}\boldsymbol{)}^{ki} \boldsymbol{\ mod\ p}$
$\boldsymbol{bk}_i$	Blinded internal i^{th} member key, or blinded intermediate key. $\boldsymbol{bk}_i = \boldsymbol{(g)}^{ki}$
$\boldsymbol{K}_{Gci}$	A key of i^{th} Cluster. $\mathbf{K}_{Gci}=(\mathbf{br}_{io})^{knci} \mathbf{mod\ p}$
$\boldsymbol{K}_G$	A key among CHs. $\mathbf{K}_G=(\mathbf{br}_{co})^{kNc} \mathbf{mod\ p}$
$\boldsymbol{h(m)}$	The digest of m

B. Overview of HSESGK

We proposed a new approach which aims to address the scalability problem while taking into consideration the dynamic aspect of the group members and dynamicity of nodes

in MANET. There are two trees on the network to avoid the robustness problem as well. Our approach is based on clustering manner. Each cluster is initiated by Cluster Head (CH), namely cluster initiator or coordinator initiator. Cluster head has then two keys; one for its cluster subgroup and another one for the interconnection among the clusters via cluster heads. Firstly, we describe our network model that is the mobile ad hoc network based on clustering that contains for example five clusters as shown in Fig. 4. There is a cluster head for each cluster and one of the cluster heads is MANET initiator or Main Cluster Head (MCH).

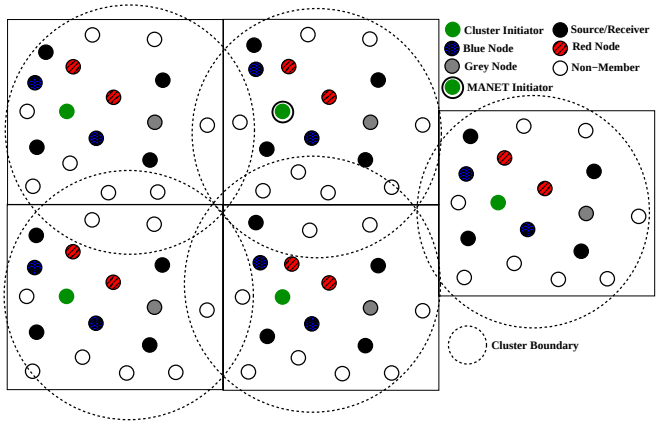


Fig. 4. MANET based on clustering.

There are many multicast routing protocols have been proposed, these protocols are classified as shown before in section 2.2. We proposed another one in the category of multicast topology, tree-based and shared tree with double trees, namely Blue tree and Red tree. All clusters then works in parallel to construct two trees. Logically, a group member views the two trees as identical trees. The group members have to be in both multicast trees.

1) Inside the Cluster: In a cluster, the cluster head (Cluster initiator) starts to initialize the process for a cluster multicast subgroup by broadcasting a join advertises message across the entire cluster. This cluster is bounded and having a fixed diameter. Each node is associated with three colors (blue, red, and grey). A node will choose its color (grey) when its total number of neighbors is less than a predefined threshold value (depending on average node degree, for instance, half of its degree). Other nodes randomly choose blue or red as their color with probability equal to 0.5. For the first received message, a grey node stores the upstream node ID and rebroadcasts the message except the node that the message is coming from. For a non-grey node, it stores the upstream node ID and rebroadcasts the message only if the upstream node is the same color, a sender/receiver, or a grey node. Based on the join response back from group members to the cluster head, two multicast trees are formed in parallel, as shown in Fig. 5. It is noted that both trees consist of group members and intermediate non-member nodes. Sure both tree are constructed in parallel and in distributer processing manner, but in blue tree‘s point of view, we find that the red’s nodes stop the broadcasting for blue tree and just blue’s nodes who broadcasting the join advertises to both blue’s nodes and grey nodes as shown in Fig. 6. As well, in red tree’s point of view, we find that the blue’s nodes stop the broadcasting for red tree and just red’s nodes who broadcasting the join advertises to both red’s nodes and grey nodes as shown in Fig. 7.

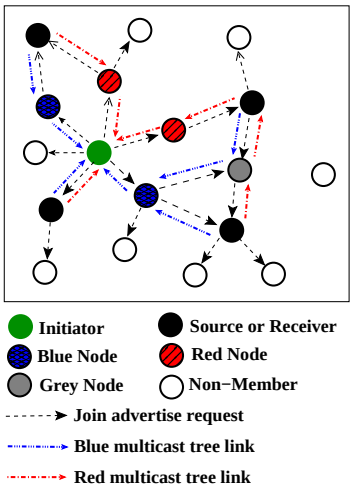


Fig. 5. Double multicast (Blue and Red) trees structure for a cluster

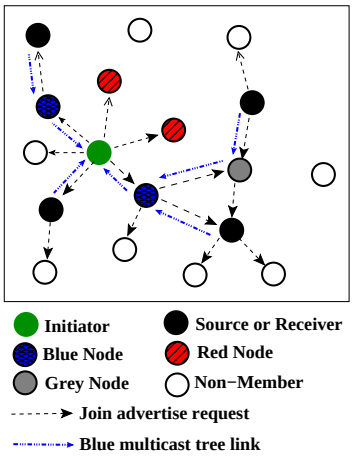


Fig. 6. Blue trees point of view for constructing itself.

2) Interconnection among the Clusters: The interconnection among the clusters is via the main cluster head (MANET initiator) starts to initialize the process for a cluster heads’ multicast subgroup by broadcasting a join advertises message across the entire MANET. We supposed the nodes no change its color, blue node still blue, red node still red, grey node still grey, and another cluster heads are source/receiver, viz, the cluster heads seems as a virtual cluster. So we can apply the same scenario that is used before in the cluster, to get both blue and red multicast trees among all cluster heads in MANET. This join advertises are broadcast across the entire

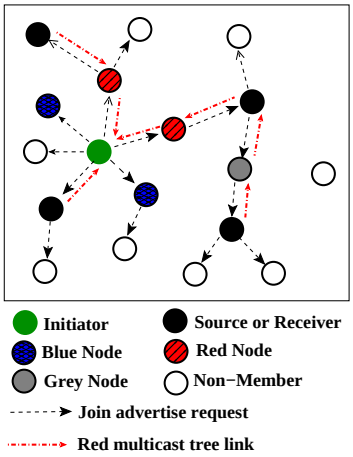


Fig. 7. Red trees point of view for constructing itself.

network as shown in Fig. 8, in which the sequence number is used to avoid the loop, and the number of hops. Based on the join response back from cluster heads to the main cluster head, two multicast trees are formed in parallel, as shown in Fig. 8. The double multicast trees among cluster heads are created and are shown in Fig. 9. Both trees consist of cluster heads, some of group members, and intermediate non-group member nodes. The resultant two trees could be disjoint or may share a common node.

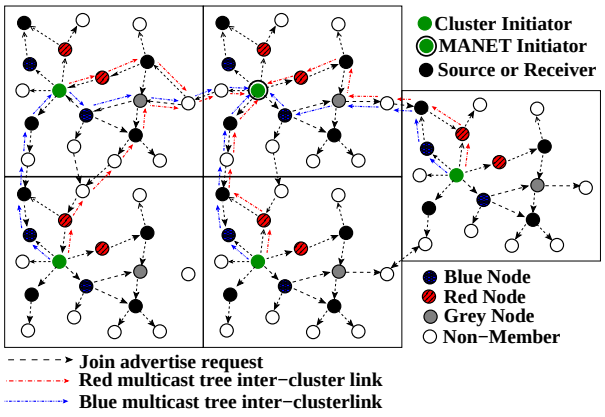


Fig. 8. Double multicast (Blue and Red) trees structure among cluster heads

As well, the double trees among cluster heads could be disjoint or may share some links in the double trees in the clusters. It is clear from the Fig. 10. Thus a dynamic double multicast trees structure for both all clusters and the subgroup of cluster heads is constructed as shown in Fig. 10. Initially the main cluster head is responsible for sending the refreshment message periodically to maintain the connection of the double trees structure. After a predefined period of time, a member could decide to act a cluster head and notify the cluster members that it is on duty to maintain the cluster subgroup. As well, a cluster head could decide to act a main cluster head

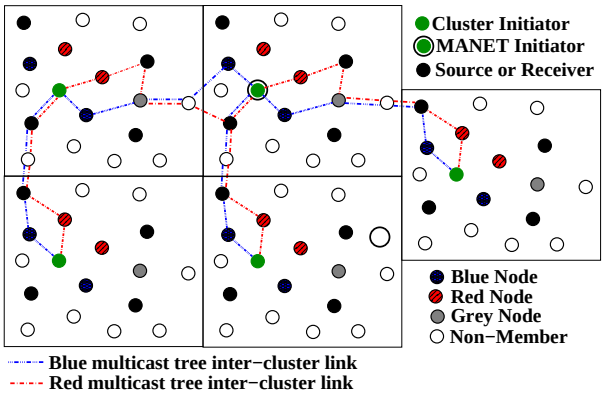


Fig. 9. Cluster Heads' multicast (Blue and Red) trees structure

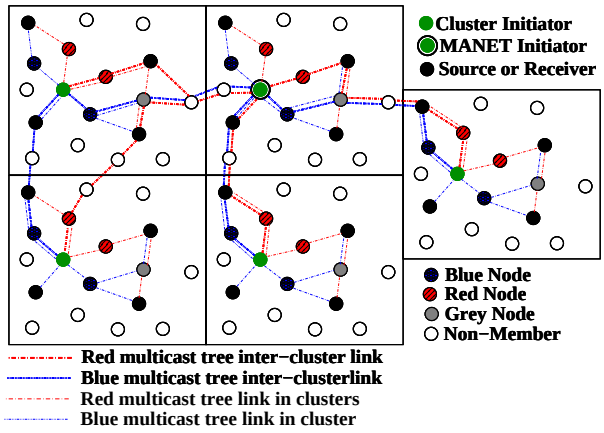


Fig. 10. Double multicast (Blue and Red) trees structure among all members in MANET

and notify the cluster heads that it is on duty to maintain the MANET group.

C. Multicast group mangement

1) A new member joins: A new member want to join a group, it could broadcast join requests to the group. The new member becomes a legitimate group member once its request is approved by any existing group member or by the cluster head of this group member. Any existing member can send replies back and send alarm "*new member*" to its cluster head. This cluster head then does the same procedure of handling join request that is similar to the above subgroup advertisement to ensure the consistency of the double multicast tree structure.

2) A member leaves: The processing of handling members who leave is more complicated than handling the joining of new members. A leaving member will not send a leaving notice. It leaves the group silently. Even if it could send a message and notify its leaving, this notice could get lost in a dynamic environment. There are a physical leaving and a logical leaving. For the physical leaving, a node moves out the range of the network or it switches its transmitter off. For a logical leaving, a node still stays inside the network, but

it does not participate in the group activity. So there are two scenarios, as follows:

First scenario: depends on detecting leaved members by its neighbors. The members are classified based in its places as follow:

1) A member is in the cluster double trees only, the neighbor of leaved member detect the leaved member and informs cluster head of its cluster to refresh the double multicast trees in this cluster.
2) A member is in cluster heads' double trees only, one of neighbor detects the leaving a member, then inform the main cluster head to refresh the double trees.
3) A member is in both a cluster double tree and cluster heads' double trees, a neighbor of leaved member detects that there is a member leaved, and inform both the main cluster head and its cluster head to refresh the double multicast trees of both cluster heads subgroup and the cluster of leaved member.

Second scenario: is based on a "*member refresh*" message that is periodically broadcasted by the cluster head across the subgroup. Each member should send an "*ack*" message back to indicate its status. The cluster head will determine whether a member remains attached or has left based on its response status within a certain time. If the cluster member on duty haven't receive "*member refresh*" message from its cluster head within a certain time, it sends a message "I am a cluster head" and send refresh the double trees in the cluster, at the same time the main cluster head detects one cluster head leaved, so it refresh the double trees of cluster heads' subgroup and so on for the main cluster head, if it leaves. This scenario is quite more costly than the first scenario but is more appropriate for a highly dynamic network like MANET where the nodes move frequently and cause the connection to be broken frequently.

D. Group key establishment protocol

The idea of subgroup key agreement protocol is that all subgroup members maintain a logic key's tree in local storage space. This key's tree is used to deduce the final common subgroup key. Our scheme is based on key's tree structure, for each subgroup; there is individual key's tree and a common subgroup key. The key's tree structure (e.g. with four members included the cluster head member, as an example) in our scheme is shown in Fig. 11.

Each member generates a private number; r_1, r_2, r_3, and r_4 for the members M_1, M_2, M_3, and M_4 respectively. The cluster head of a subgroup generates the numbers r and r_0, and informs all other members in its subgroup.The two numbers (r, r_0) at the two ends of the key tree for efficient group key refreshing and the cluster head role switching. Also, it is responsible for handling the member join and leave. All members reply its cluster head by intermediate keys to calculating keys. In this example: a subgroup contains four nodes. The cluster head multicast the intermediated blind keys to all members. So, each member deduces locally the final common subgroup key. The given parameters' value for each node: g=2, p = 13, r=3 then br = g^r mod p = 2^3 mod 13 = 8,

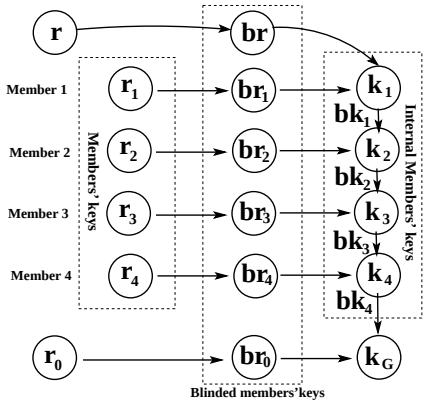


Fig. 11. Key's tree structure to generate group key (K_G) with 4 members

r_0 =5 then $br_0 = g^{r_0}$ mod p = 2^5 mod 13 = 6. Each member i, $\forall i \epsilon [1,4]$, can calculate the K_G as follows:

Inside M_1
r_1 = 4, $br_1 = g^{r_1}$ mod p = 2^4 mod 13= 3,
$k_1 = {br_1}^r$ mod p= 3^3 mod 13 =1,
$bk_1 = g^{k_1} = 2^1$ =2
$\Longrightarrow k_1 = {br}^{r_1}$ **mod p** = 8^4 **mod 13 = 1**
$\Longrightarrow k_2 = {br_2}^{k_1}$ mod p = 6^1 mod 13 = 6
$\Longrightarrow k_3 = {br_3}^{k_2}$ mod p = 11^6 mod 13 = 12
$\Longrightarrow k_4 = {br_4}^{k_3}$ mod p = 12^{12} mod 13 = 1
$\Longrightarrow K_G = {br_0}^{k_4}$ **mod p** = 6^1 **mod 13 = 6**

Inside M_2
r_2 = 5, $br_2 = g^{r_2}$ mod p = 2^5 mod 13= 6,
$k_2 = {br_2}^{k_1}$ mod p= 6^1 mod 13 =6,
$bk_2 = g^{k_2} = 2^6$ =64
$\Longrightarrow k_2 = {bk_1}^{r_2}$ **mod p** = 2^5 **mod 13 = 6**
$\Longrightarrow k_3 = {br_3}^{k_2}$ mod p = 11^6 mod 13 = 12
$\Longrightarrow k_4 = {br_4}^{k_3}$ mod p = 12^{12} mod 13 = 1
$\Longrightarrow K_G = {br_0}^{k_4}$ **mod p** = 6^1 **mod 13 = 6**

Inside M_3
r_3 = 7, $br_3 = g^{r_3}$ mod p = 2^7 mod 13= 11,
$k_3 = {br_3}^{k_2}$ mod p= 11^6 mod 13 =12,
$bk_3 = g^{k_3} = 2^{12}$ =4096
$\Longrightarrow k_3 = {bk_2}^{r_3}$ **mod p** = 64^7 **mod 13 = 12**
$\Longrightarrow k_4 = {br_4}^{k_3}$ mod p = 12^{12} mod 13 = 1
$\Longrightarrow K_G = {br_0}^{k_4}$ **mod p** = 6^1 **mod 13 = 6**

Inside M_4
r_4 = 6, $br_4 = g^{r_4}$ mod p = 2^6 mod 13= 12,
$k_4 = {br_4}^{k_3}$ mod p= 12^{12} mod 13 =1,
$bk_4 = g^{k_4} = 2^1$ =2
$\Longrightarrow k_4 = {bk_4}^{r_4}$ **mod p** = 12^{12} **mod 13 = 1**
$\Longrightarrow K_G = {br_0}^{k_4}$ **mod p** = 6^1 **mod 13 = 6**

1) Initialization: CH announces its role and broadcasts two random keys (r, r_0) and its br_c, br, and br_0. Each member has unique identifier (ID) that is given by its cluster head when joining the group. At the initialization phase, the members are sorted by their ID. M_i, $\forall$ i ϵ [1,N_c], (where N_c is number of subgroup's members) generates a private random number r_i then compute the br_i and send it to its CH. CH is then responsible for computing $k_1 \ldots k_{N_C}$ and $bk_1 \ldots bk_{N_c}$ and

then multicasts them to the subgroup's members.

All keying materials are put in one package and the order of blinded intermediate key materials shows the structure of the key tree. Each member can thus deduce the common subgroup key (K_G). The time diagram of initialization process to deduce the common group key (K_G) in a subgroup is shown in Fig. 12 for each cluster(i.e.either members' clusters or CH's cluster).

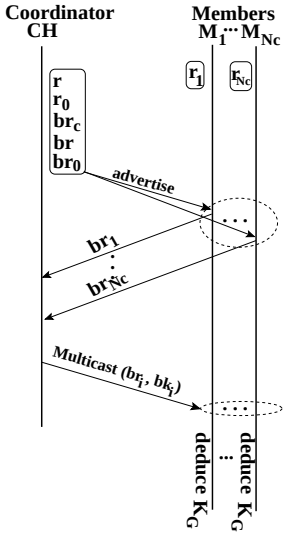
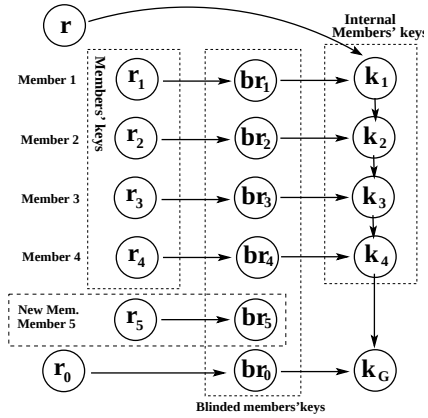


Fig. 12. Time diagram of initialization process of deducing group key (K_G) in a subgroup

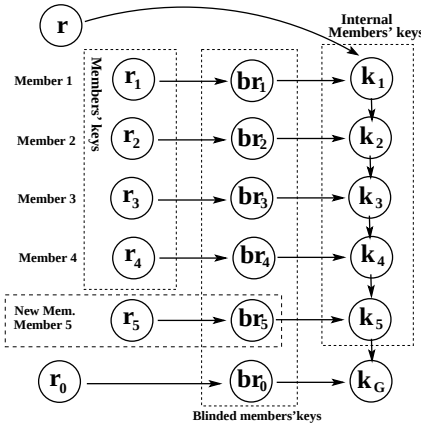
2) Member join: A new member can be easily added into the nearest cluster as described before in section III-C1. The double trees are constructed. The cluster head insert the new member in the current rightmost position and give it an ID. The cluster head does not generate any random key but still provides key independence. Given blinded keys, the new member deduce the new common subgroup key, however it cannot deduce the previous common subgroup key.

Fig. 13 depicts Key tree structure to generate group key (K_G), while a new member wants to join a subgroup. We take the same example used before in this section with adding a new member M_5. The given parameters' value for each member: g=2, p = 13, r=3 then br = g^r mod p = 2^3 mod 13 = 8, r_0 =5 then $br_0 = g^{r_0}$ mod p = 2^5 mod 13 = 6. Each member i, $\forall$ i ϵ [1, 5], can calculate the K_G as follows:

Inside M_1
r_1 = 4, $br_1 = g^{r_1}$ mod p = 2^4 mod 13= 3,
$k_1 = {br_1}^r$ mod p= 3^3 mod 13 =1,
$bk_1 = g^{k_1} = 2^1$ =2
$\Longrightarrow k_1 = {br}^{r_1}$ **mod p** = 8^4 **mod 13 = 1**
$\Longrightarrow k_2 = {br_2}^{k_1}$ mod p = 6^1 mod 13 = 6
$\Longrightarrow k_3 = {br_3}^{k_2}$ mod p = 11^6 mod 13 = 12
$\Longrightarrow k_4 = {br_4}^{k_3}$ mod p = 12^{12} mod 13 = 1
$\Longrightarrow k_5 = {br_5}^{k_4}$ mod P = 3^1 mod 13 = 3
$\Longrightarrow K_G = {br_0}^{k_5}$ **mod p** = 6^3 **mod 13 = 8**



(a) Before joining a member 5



(b) After joining a member 5

Fig. 13. Key tree structure to generate group key (K_G), while a member join a subgroup

Inside M_2
r_2 = 5, $br_2 = g^{r_2}$ mod p = 2^5 mod 13= 6,
$k_2 = {br_2}^{k_1}$ mod p= 6^1 mod 13 =6,
$bk_2 = g^{k_2} = 2^6$ =64
$\Longrightarrow k_2 = {bk_1}^{r_2}$ **mod p** = 2^5 **mod 13 = 6**
$\Longrightarrow k_3 = {br_3}^{k_2}$ mod p = 11^6 mod 13 = 12
$\Longrightarrow k_4 = {br_4}^{k_3}$ mod p = 12^{12} mod 13 = 1
$\Longrightarrow k_5 = {br_5}^{k_4}$ mod P = 3^1 mod 13 = 3
$\Longrightarrow K_G = {br_0}^{k_5}$ **mod p** = 6^3 **mod 13 = 8**

Inside M_3
r_3 = 7, $br_3 = g^{r_3}$ mod p = 2^7 mod 13= 11,
$k_3 = {br_3}^{k_2}$ mod p= 11^6 mod 13 =12,
$bk_3 = g^{k_3} = 2^{12}$ =4096
$\Longrightarrow k_3 = {bk_2}^{r_3}$ **mod p** = 64^7 **mod 13 = 12**
$\Longrightarrow k_4 = {br_4}^{k_3}$ mod p = 12^{12} mod 13 = 1
$\Longrightarrow k_5 = {br_5}^{k_4}$ mod P = 3^1 mod 13 = 3
$\Longrightarrow K_G = {br_0}^{k_5}$ **mod p** = 6^3 **mod 13 = 8**

Inside M_4
r_4 = 6, br_4= g^{r_4} mod p = 2^6 mod 13= 12,
k_4 = ${br_4}^{k_3}$ mod p= 12^{12} mod 13 =1,
bk_4 = g^{k_4} = 2^1 =2
$\Longrightarrow$ k_4 = ${bk_3}^{r_4}$ **mod p** = 4096^6 **mod 13 = 1**
$\Longrightarrow$ k_5 = ${br_5}^{k_4}$ mod P = 3^1 mod 13 = 3
$\Longrightarrow$ K_G = ${br_0}^{k_5}$ **mod p** = 6^3 **mod 13 = 8**

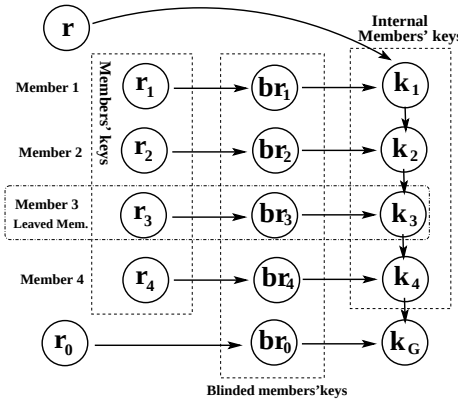
Inside M_5
r_5 = 4, br_5= g^{r_4} mod p = 2^4 mod 13= 3,
k_5 = ${br_5}^{k_3}$ mod p= 3^1 mod 13 =3,
bk_5 = g^{k_5} = 2^3 =8
$\Longrightarrow$ k_5 = ${bk_4}^{r_5}$ **mod P** = 2^4 **mod 13 = 3**
$\Longrightarrow$ K_G = ${br_0}^{k_5}$ **mod p** = 6^3 **mod 13 = 8**

3) Member leave: A member can be easily leaved from its cluster as described before in section III-C2. The double trees are constructed. It is possible that the leaved member is either a member in a cluster (subgroup) or a cluster head. *Case 1*: leaving of a member in a cluster, its cluster head generates a new random key r' instead of r and multicast the blinded value br' as well as other intermediate blinded keys. Each member i, $\forall$ i ϵ [1, N_c]\{leaved member}, can then calculate the K_{G_c}. *Case 2*: leaving of cluster head, a cluster member on duty acts as a cluster head as before, moreover, the main cluster head detects a cluster head leaved, so the leaved process seems like two leaved members (but really one leaved member), one from a cluster's subgroup and another from the cluster heads' subgroup. In two cases, the leaved process simply takes place in a subgroup as shown in Fig. 14, that depicts key tree structure to generate both group key (K_{G_c}) for the cluster of leaved member and group key (K_G) for cluster heads' subgroup via the same process, while a member leaves the multicast group.

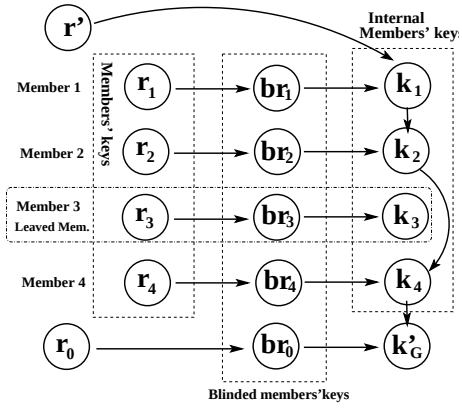
Also, we take the same example used before in this section with leaving a member M_3 in *Case 1*. The given parameters' value for each member: g=2, p = 13, r'=5 then br' = $g^{r'}$ mod p = 2^5 mod 13 = 6, r_0 =5 then br_0 = g^{r_0} mod p = 2^5 mod 13 = 6. Each member i, $\forall$ i ϵ [1, 5]\{3}, can calculate the K_G as follows:

Inside M_1
r_1 = 4, br_1= g^{r_1} mod p = 2^4 mod 13= 3,
k_1 = ${br_1}^{r}$ mod p= 3^5 mod 13 =9,
bk_1 = g^{k_1} = 2^9 =512
$\Longrightarrow$ k_1 = ${br'}^{r_1}$ **mod p** = 6^4 **mod 13 = 9**
$\Longrightarrow$ k_2 = ${br_2}^{k_1}$ mod p= 6^9 mod 13 = 5
$\Longrightarrow$ k_4 = ${br_4}^{k_2}$ mod p= 12^5 mod 13 = 12
$\Longrightarrow$ K_G = ${br_0}^{k_4}$ **mod p**= 6^{12} **mod 13 = 1**

Inside M_2
r_2 = 5, br_2= g^{r_2} mod p = 2^5 mod 13= 6,
k_2 = ${br_2}^{k_1}$ mod p= 6^9 mod 13 =5,
bk_2 = g^{k_2} = 2^5 =32
$\Longrightarrow$ k_2 = ${bk_1}^{r_2}$ **mod p**= 512^5 **mod 13 = 5**
$\Longrightarrow$ k_4 = ${br_4}^{k_2}$ mod p= 12^5 mod 13 = 12
$\Longrightarrow$ K_G = ${br_0}^{k_4}$ **mod p**= 6^{12} **mod 13 = 1**



(a) Before leaving a member 3



(b) After leaving a member 3

Fig. 14. Key tree structure to generate group key (K_G), while a member leaves the member group

Inside M_4
r_4 = 6, br_4= g^{r_4} mod p = 2^6 mod 13= 12,
k_4 = ${br_4}^{k_2}$ mod p= 12^5 mod 13 =12,
bk_4 = g^{k_4} = 2^{12} =4096
$\Longrightarrow$ k_4 = ${bk_2}^{4_4}$ mod p= 32^6 mod 13 = 12
$\Longrightarrow$ K_G = ${br_0}^{k_4}$ **mod p**= 6^{12} **mod 13 = 1**

4) Group key refresh/reinforce: The group key may need to be changed periodically, and may not be related to any change of group membership. The purpose of refreshing the group key periodically is to prevent the long time use of group keys which could be compromised. This process can be implicitly done during the switch of cluster head, or explicitly performed by the cluster head which generates a new random key r" and multicasts the blinded value br" as well as other intermediate blinded keys. Then each member i, $\forall$ i ϵ [1, N_c], can calculate the K_{G_c} as described in section III-D1. Refresh/reinforce process take place independently in each cluster, as well in the cluster heads' subgroup. That decreases the traffic control overheads and increases the scalability in MANET.

IV. Discussion

The goal of all these protocols include such as minimal control overhead, minimal processing overhead, multi-hop routing

capability, dynamic topology maintenance, loop prevention, or more secure. However many multicast routing protocols don't perform well in MANETs because in a highly dynamic environment, node move arbitrarily, and man-in-middle problem. Our paper focuses on the key management schemes that are important part of the security. So key management is an essential cryptographic primitive upon which other security primitives such as privacy, authenticity and integrity are built. As well, it has to be satisfied some features such as *Security*, *Reliability*, *Scalability*, *Robustness*, and *power consumption*, as follows:

Security: intrusion tolerance means system security should not succumb to a single, or a few, compromised nodes. So, the key management schemes should ensure no unauthorized node receives key material that can later be used to prove status of a legitimate member of the network. Here the key is computed in distributed manner, and the member provides a trusted group communication. Other issues are trust management, vulnerability. Also, proper key lengths and cryptographic algorithms of adequate strength are assumed.

Reliability: depends on the key distribution, storage and maintenance and make sure that keys are properly distributed among the nodes, safely stored where intruders aren't able to hack the keys and should be properly maintained. In our proposed, each member can deduce the common group key depending on a private value, not be exchanged and some common parameters shared among members. It means that no need to exchange the group key, so this group key is stored locally on a member with a certain security manner.

Scalability: the key management operations should finish in a timely manner despite a varying number of nodes and node densit ies. It makes use the occupied network bandwidth of network management traffic as low as possible to increase nodes' density. Making use of clustering scheme, decreases the control overhead traffic due to the double trees creation, and increase the number of members in the MANET with lowest control overhead.

Robustness: the key management system should survive despite Denial-of-Service (DoS) attacks and unavailable nodes. Because of dynamicity of the group members, necessary key management operation should execute in a timely manner, in order not to make a isolated partition in the network. In our proposal, multiple trees are used for the robustness and avoid fault tolerance.

Power consumption: Energy saving, despite recent advances in extending battery life, is still an important issue. Basically, MANETs protocols must be aware that a mobile node has a finite battery capacity. In another side, decreases the processing time, as low as possible to increase the life time of nodes. We believe that delay and delay jitter should be given the highest priority when dealing with for example video traffic over the wireless network. It means that many researchers have focused and emphasized on saving power of the node battery to last for longer time without recharging as mentioned in [58].

V. Conclusion

MANET is one of the most important and unique applications. Due to the nature of unreliable wireless medium data transfer is a major problem in MANET and it lacks security and reliability of data. A Key management is vital part of security. Key management protocols then play a key role in any secure group communication architecture. Moreover in MANET, members can join and leave the group dynamically during the whole session, plus the nodes movement. So, the key management is an important challenge because of its dynamism that affects considerably its performance. In this paper, we have studied the different key management schemes for MANET and proposed a new scheme namely HSESGK, which is an efficient/scalable hierarchical key management scheme for MANET multicast. In our scheme, the group members deduce the group key in a distributed manner. This hierarchical contains two levels only, first level for all clusters' heads as a main group's members; the second level for all clusters' members. Then there is a secret key obtained in a distributed manner for each cluster subgroup, and another secret key for clusters' heads subgroup. It is shown that our scheme reduces significantly the overall security overhead of member's join or leave compared to all other schemes and more reducing the ratio between control overheads and data. it is satisfied for some features such as Security, Reliability, Scalability, Robustness, and power consumption.

References

[1] M. Younis and S. Z. Ozer, "Wireless ad hoc networks: technologies and challenges," *Wireless Communications and Mobile Computing*, vol. 6, no. 7, pp. 889–892, 2006.

[2] S. Guo and O. W. W. Yang, "Energy-aware multicasting in wireless ad hoc networks: A survey and discussion," *Computer Communications*, vol. 30, no. 9, pp. 2129–2148, 2007.

[3] J. Wang, C. Wang, and Q. Wu, Eds., *Ad Hoc Mobile Wireless Network*. Beijing, National defense industry press, 2004.

[4] C. Xiao and W. Jie, "Multicasting techniques in mobile ad hoc networks," in *The handbook of ad hoc wireless networks*, I. Mohammad and C. D. Richard, Eds. CRC Press, Inc., 2003, pp. 25–40, 989714.

[5] L. Junhai, Y. Danxia, and . all, "Research on routing security in manet," *Application Research of Computers*, vol. 25, no. 1, pp. 243–245, 2008.

[6] R. A. and K. C. Shet, "Hierarchical approach for key management in mobile ad hoc networks," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 5, no. 1, pp. 87–95, 2009.

[7] M.-S. Bouassida, I. Chrisment, and O. Festor, "Group key management in manets," *International Journal of Network Security (IJNS)*, vol. 6, no. 1, pp. 67–79, 2008.

[8] W. Diffie and M. E. Hellman, "New directions in cryptography," *Information Theory, IEEE Transactions on*, vol. 22, no. 6, pp. 644–654, 1976.

[9] L. Harn, M. Mehta, and H. Wen-Jung, "Integrating diffie-hellman key exchange into the digital signature algorithm (dsa)," *Communications Letters, IEEE*, vol. 8, no. 3, pp. 198–200, 2004.

[10] R. C. W. Phan, "Fixing the integrated diffie-hellman-dsa key exchange protocol," *Communications Letters, IEEE*, vol. 9, no. 6, pp. 570–572, 2005.

[11] M. Francis, M. Sangeetha, and A. Sabari, "A survey of key management technique for secure and reliable data transmission in manet," *International Journal of Advanced Research in Computer Science and Software Engineering (IJAARCSSE)*, vol. 3, no. 1, pp. 22–27, 2013.

[12] K. Hussain, A. H. Abdullah, S. Iqbal, K. Awan, and F. Ahsan, "Efficient cluster head selection algorithm for manet," *Journal of Computer Networks and Communications*, vol. 2013, no. 7, pp. 1–7, 2013.

[13] P. Sivaprakasam and R. Gunavathi, "An efficient clusterhead election algorithm based on maximum weight for manet," in *Advanced Computing (ICoAC), 2011 Third International Conference on*, Dec 2011, pp. 315–320.

[14] S. Mehta, P. Sharma, and K. Kotecha, "A survey on various cluster head election algorithms for manet," in *Engineering (NUiCONE), 2011 Nirma University International Conference on*, Dec 2011, pp. 1–6.

[15] D. Hongmei, W. Li, and D. P. Agrawal, "Routing security in wireless ad hoc networks," *Communications Magazine, IEEE*, vol. 40, no. 10, pp. 70–75, 2002.

[16] L. Abusalah, A. Khokhar, and M. Guizani, "A survey of secure mobile ad hoc routing protocols," *Communications Surveys & Tutorials, IEEE*, vol. 10, no. 4, pp. 78–93, 2008.

[17] D. B. Johnsort, "Routing in ad hoc networks of mobile hosts," in *Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on*, 1994, pp. 158–163.

[18] L. Wenjing, L. Wei, and F. Yuguang, "Spread: enhancing data confidentiality in mobile ad hoc networks," in *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, vol. 4, 2004, pp. 2404–2413 vol.4.

[19] R. Hauser, M. Consulting, T. Przygienda, and G. Tsudik, "Lowering security overhead in link state routing," *Computer Networks*, vol. 31, no. 8, pp. 885–894, 1999.

[20] H. Yih-Chun, A. Perrig, and D. B. Johnson, "Packet leashes: a defense against wormhole attacks in wireless networks," in *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, vol. 3, 2003, pp. 1976–1986 vol.3.

[21] B. Sonja and B. Jean-Yves Le, "Performance analysis of the confidant protocol," 2002, 513828 226-236.

[22] V. D. Park and M. S. Corson, "A highly adaptive distributed routing algorithm for mobile wireless networks," in *INFOCOM '97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Driving the Information Revolution., Proceedings IEEE*, vol. 3, 1997, pp. 1405–1413 vol.3.

[23] M. Sergio, T. J. Giuli, L. Kevin, and B. Mary, "Mitigating routing misbehavior in mobile ad hoc networks," 2000, 345955 255-265.

[24] J. J. Garcia-Luna-Aceves and M. Spohn, "Source-tree routing in wireless networks," in *Network Protocols, 1999. (ICNP '99) Proceedings. Seventh International Conference on*, 1999, pp. 273–282.

[25] W. Yong, G. Attebury, and B. Ramamurthy, "A survey of security issues in wireless sensor networks," *Communications Surveys & Tutorials, IEEE*, vol. 8, no. 2, pp. 2–23, 2006.

[26] P. Papadimitratos and Z. J. Haas, "Secure routing: Secure data transmission in mobile ad hoc networks," in *ACM Workshop on Wireless Security (WiSe 2003)*, vol. 1-58113-585-8/02/0009, San Diego, California, USA, 2003, pp. 41–50.

[27] L. Lilien, "Developing pervasive trust paradigm for authentication and authorization," in *Cracow Grid Workshop*. Institute of Computer Science, AGH University of Science and Technology, Cracow,Poland: Academic Computer Centre CYFRONET AGH, 2004, pp. 42–49.

[28] P. Jacquet, P. Muhlethaler, and A. Qayyum, "Optimized link state routing protocol," in *RFC 3626*, 2003.

[29] C. E. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing," in *Mobile Computing Systems and Applications, 1999. Proceedings. WMCSA '99. Second IEEE Workshop on*, 1999, pp. 90–100.

[30] T. H. Clausen, G. Hansen, L. Christensen, and G. Behrmann, "The optimized link state routing protocol evaluation through experiments and simulation," in *In Proceedings of the IEEE Symposium on Wireless Personal Mobile Communications*, Mindpass Center for Distributed Systems, Aalborg University, Fredrik Bajers Vej 7E, DK-9220 Aalborg, Denmark, 2001.

[31] Z. Manel Guerrero, "Secure ad hoc on-demand distance vector routing," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 6, no. 3, pp. 106–107, 2002, 581312.

[32] H. Yih-Chun and B. J. David, "Securing quality-of-service route discovery in on-demand routing for ad hoc networks," 2004, 1029120 106-117.

[33] X. Chen and J. Wu, "Multicasting techniques in mobile ad-hoc networks," *The Handbook of Ad-hoc Wireless Networks*, pp. 25–40, 2003.

[34] T. P. Singh, Neha, and V. Das, "Multicast routing protocols in manets," *International Journal of Advanced Research in Computer Science and Software Engineering (IJAARCSSE)*, vol. 2, no. 1, pp. 1–6, 2012.

[35] J. Luo, D. Ye, L. Xue, and M. Fan, "A survey of multicast routing protocols for mobile ad-hoc networks," *Communications Surveys & Tutorials, IEEE*, vol. 11, no. 1, pp. 78–91, 2009.

[36] N. Meghanathan, "Survey of topology-based multicast routing protocols for mobile ad hoc networks," *International Journal of Communication Networks and Information Security (IJCNIS)*, vol. 3, no. 2, pp. 124–137, 2011.

[37] L. Junhai, X. Liu, and Y. Danxia, "Research on multicast routing protocols for mobile ad-hoc networks," *Computer Networks*, vol. 52, no. 5, pp. 988–997, 2008.

[38] C. Siva, R. Murthy, and B. S. Manoj, *Ad Hoc Wireless Networks Architectures and Protocols*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2004.

[39] C. K. Toh, *Ad Hoc Wireless Networks: Protocols and Systems*. 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.

[40] A. C. F. Chan, "Distributed symmetric key management for mobile ad hoc networks," in *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, vol. 4, 2004, pp. 2414–2424 vol.4.

[41] B. Aziz, E. Nourdine, and E. K. Mohamed, "A recent survey on key management schemes in manet," in *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on*, 2008, pp. 1–6.

[42] R. Anderson, C. Haowen, and A. Perrig, "Key infection: smart trust for smart dust," in *Network Protocols, 2004. ICNP 2004. Proceedings of the 12th IEEE International Conference on*, 2004, pp. 206–215.

[43] V. Gerardo del, G. Roberto, C. mez, and rdenas, "Overview the key management in ad hoc networks," 2005, 2153214 397-406.

[44] W. Bing, W. Jie, B. F. Eduardo, I. Mohammad, and M. Spyros, "Secure and efficient key management in mobile ad hoc networks," *J. Netw. Comput. Appl.*, vol. 30, no. 3, pp. 937–954, 2007, 1231774.

[45] AnilKapil and SanjeevRana, "Identity-based key management in manets using public key cryptography," *International journal of Security (IJS)*, vol. 3, no. 1, pp. 1–26, 2009.

[46] Y. D. Bing Wu, Jie Wu, "An efficient group key management scheme for mobile ad hoc networks," *International Journal of Security and Networks (IJSN)*, vol. 4, no. 2, pp. 125–134, 2009.

[47] D. B. Shacham, X. Boyen, and Hovav, "Short group signatures," in *In Advances in CryptologyCrypto04, Lecture Notes in Computer Science, vol. 3152*, vol. 3152, 2004, pp. 41–55.

[48] R. PushpaLakshmi and A. V. A. Kumar, "Cluster based composite key management in mobile ad hoc networks," *International Journal of Computer Applications*, vol. 4, no. 7, pp. 30–35, 2010.

[49] X. Hai-tao, "A cluster-based key management scheme for manet," in *Intelligent Systems and Applications (ISA), 2011 3rd International Workshop on*, May 2011, pp. 1–4.

[50] ThairKhdour and A. Aref, "A hybrid schema zone-based key management for manets," *Journal of Theoritical and Applied Information Tecnology (JATIT)*, vol. 35, no. 2, pp. 175–183, 2012.

[51] C.-c. Chiang, H.-K. Wu, W. Liu, and M. Gerla, "Routing in clustered multihop, mobile wireless networks with fading channel," in *in Proceedings of IEEE SICON*, 1997, pp. 197–211.

[52] D. Gavalas, G. Pantziou, C. Konstantopoulos, and B. Mamalis, "Clustering of mobile ad hoc networks: An adaptive broadcast period approach," in *Communications, 2006. ICC '06. IEEE International Conference on*, vol. 9, June 2006, pp. 4034–4039.

[53] R. Selvam and V. Palanisamy, "An optimized cluster based approach for multi-source multicast routing protocol in mobile ad hoc networks with differential evolution," in *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on*, March 2012, pp. 115–120.

[54] C. Bemmoussat, F. Didi, and M. Feham, "Cluster based routing protocol in wireless mesh network," in *Computer Applications Technology (ICCAT), 2013 International Conference on*, Jan 2013, pp. 1–6.

[55] A. EL-SAYED, "Clustering based group key management for manet," in *International Conference on Advances in Security of Information and Communication Networks (SecNet'2013), 3-5 September, 2013, Cairo, Egypt.*, vol. 382, 2013, pp. 11–26.

[56] W. Wei and A. Zakhor, "Multiple tree video multicast over wireless ad hoc networks," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 1, pp. 2–15, 2007.

[57] B. R. Tamma, A. Badam, C. Siva Ram Murthy, and R. R. Rao, "K-tree: A multiple tree video multicast protocol for ad hoc wireless networks," *Computer Networks*, vol. 54, no. 11, pp. 1864–1884, 2010.

[58] H. M. Asif, T. R. Sheltami, and E. E. Shakshuki, "Power consumption optimization and delay minimization in manet," in *Proceedings of the 6th International Conference on Advances in Mobile Computing and Multimedia*, ser. MoMM '08. New York, NY, USA: ACM, 2008, pp. 67–73. [Online]. Available: http://doi.acm.org/10.1145/1497185.1497202

Ayman EL-SAYED received the BSc degree in computer science and engineering in 1994, the masters degree in computer networks in 2000 from the University of Menoufiya, Egypt, and the PhD degree in computer network in 2004 from Institute National De Polytechnique De Grenoble INPG, France. He is an associate professor in the Computer Science and Engineering Department, Faculty of Electronic Engineering, Menoufiya University, Egypt. He is specialized in soft computing, algorithms, and data structure. Also, his interests include multicast routing protocols, application-level multicast techniques, multicast on both mobile network and mobile IP, and image processing techniques. Also, there are other interesting topics such as bioinformatics, biocomputing, and bio computer. He is an approved supervisor for MSc and PhD programs in various University. He has completed various project in government and private organization. He has published more than 45 research papers in international Journals and two books about OSPF protocol and multicast protocols. Currently, he is serving as an editorial board member in various international Journals and conferences. He is a senior member of the IEEE.

Human Recognition System using Cepstral Information

Emna RABHI
Département Génie Electrique
Université Tunis EL Manar, Ecole Nationale d'Ingénieurs de TUNIS, LR Signal, Image et Technologies de l'information, BP 37, Belvédère, 1002
Tunis, Tunisie
emna.rabhi@gmail.com

Zied Lachiri
Département Génie physisque et instrumentation
Institut National des sciences appliquées et de technologie
Tunis, Tunisie
zied.lachiri@enit.rnu.tn

***Abstract*— This paper presents a new method for human recognition using the cepstral information. The proposed method consists in extracting the Linear Frequency Cepstral Coefficients (LFCC) from each heartbeat in the homomorphic domain. Thus, the Hidden Markov Model (HMM) under Hidden Markov Model Toolkit (HTK) is used for electrocardiogram (ECG) classification. To evaluate the performance of the classifier, the number of coefficients and the number of frequency bands are varied. Concerning the HMM topology, the number of Gaussians and states are also varied. The best rate is obtained with 32 coefficients, 24 frequency bands, 1 Gaussian and 5 states. Further, the method is improved by adding dynamic features: the first order delta (Δ) and energy (E) to the coefficients.**
The approach is evaluated on 18 healthy signals of the MIT_BIH database.
The obtained results reveal which LFCC with energy that make a 33 dimensional feature vector leads to the best human recognition rate which is 99.33%.

***Keywords*— Electrocardiogram(ECG); Linear Frequency Cepstral Coefficients(LFCC); Hidden Markov Model (HMM).**

I. INTRODUCTION

Biometrics is a secure alternative to traditional methods of identity verification of individuals, such as passwords. The increasing need for security leads to the growth of biometrics, and the search for new biometric technology becomes topical. Biometrics uses the physiological or behavioral characteristics that are unique to each one in order to determine the identity of individuals [1]. There are several types of biometric methods such as the fingerprint, hand geometry, face recognition, iris, etc… Unfortunately, with the development of technology falsification, these features can be forged. Hence, the need for new research which could be difficult to imitate is behind the use of the physiological signals ECG as a biometric characteristic [9].

Biel et al. [4] are the earliest researchers who have worked with ECG as a biometric characteristic. They have extracted twelve features from each record for human recognition in the time domain. For classification, the SIMCA (Soft Independent Modeling of Class Analogy) model was used.

Shen et al. [13] have limited their interest in their recognition algorithm, to only a few fiducial points, surrounding the QRS complex. They have extracted seven features. To evaluate the performance of their method, they have used neural network (DBNN) and template matching.

Israel et al. [10] have determined the ECG signal peaks in the time domain by finding local maxima in the regions surrounding each of the P, R and T waves. Then, 15 features were extracted which denote the time distance between detected features.

To sum up, the major shortcomings of the previous works is the following: the method in detection of fiducial points; because no universally acknowledged definition exists for defining exactly where the wave boundaries lie and the physiological changes of the heart.

In this paper, an attempt is made to present a new method based on features extraction of ECG without fiducial detection. To improve the identification accuracy, an approach based on cepstral information is introduced. It is reflected in the compute of the LFCC from each heart beat of the signal ECG.

This paper is organized as follows. Section 2 presents the human Recognition System that contains a description of the features extraction and the classifier HMM. Section 3 is about the experimental results of our method. Finally, conclusions are presented in Section 4.

II. HUMAN RECOGNITION SYSTEM

The electrocardiogram is a non-periodic [3] but highly recurrent signal that is why we used the cepstral coefficients with linear filter banks [6,7]. The human recognition system consists in extraction of LFCC of each heart beat which were varied from 12 to 34. A package named LFCC-RASTAMAT [15] was adapted for extracting LFCC features.

In order to improve the results, dynamic features were added to the coefficients which are: first order delta (Δ) and energy (E).

In Electrocardiogram classification, Hidden Markov Model (HMM) was used. To choose the best model topology of the classifier, the number of states was varied from 5 to 7 and the number of Gaussians from 1 to 5.

A. Features Extraction

- **Linear Frequency Cepstral Coefficients (LFCC)**

The cepstral coefficients are calculated from the modulus

of the fast Fourier transform of the signal windowed by Hamming. An analysis filter bank module converts energy by bands coefficients. The scale filter banks used in this work is linear to calculate the Linear Frequency Cepstral Coefficients (LFCC) [14].

The block diagram illustrating the steps of calculating LFCC is as follows:

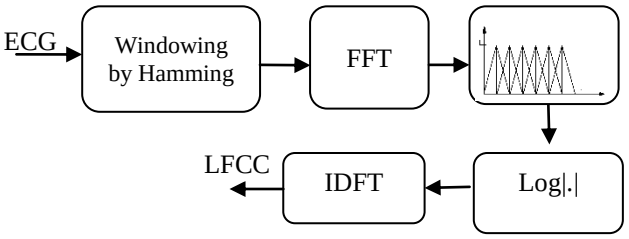


Fig 1. LFCC Algorithm.

- **Energy (E)**

In order to identify the variations of an ECG signal, the energy is often evaluated in several successive frames. For a frame "i", the energy parameter is computed as follows:

$$E = \sum_{n=0}^{N-1} X_i^2[n] \quad (1)$$

- **The first order delta (Δ)**

Calculating the first order delta of an ECG signal means computing its first derivative. The first derivative is the rate of change of y with x: dy/dx note that x and y denote the coordinates of signal's points. In other words, it is the slope of the tangent to the signal at each point.

B. Hidden Markov Model (HMM)

A HMM is a stochastic model. After learning, it is able to provide the probability of generating a given set of observations. This model is defined by a sequence of N states qt connected by transitions aij at time t [5].

The training of HMM model according to the prototype model is initialized by the Viterbi algorithm, and then re-estimated by the Baum-Welch algorithm. The extracted parameters are parameterized vectors whose representation defined the set of observations O={o1,o2,…,oT} which are useful in the process of modeling the HMM.

The Bakis Model is adopted as a HMM topology. It is a model left to right with 5 states whose initial and final states do not emit vectors of observations as shown in Figure2. The mean vector of the used model is initialized to zero and a variance vector is initialized to one.

For building and manipulating the Hidden Markov models, a platform Hidden Markov Model Toolkit (HTK) [11] was used. It is a platform formed by a set of libraries and tools software written in language C, its modularity allows to develop its own software and uses the tools available directly [12].

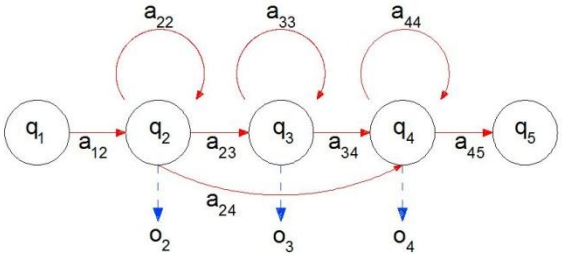


Fig 2. Bakis Model.

- **Training**

Recognition of data requires the training of parameters associated with their respective HMM from a training corpus. The estimation of model parameters is made by applying the Baum-Welch optimal algorithm until convergence and re-estimate probabilities of emission and transition.

- **Recognition**

Recognition is achieved by the Viterbi algorithm, it is used to find the most likely sequence of hidden states, called the Viterbi path, which results in a sequence of observed events.

The different steps of the recognition algorithm construction of HMM under platform HTK are illustrated in Figure 3.

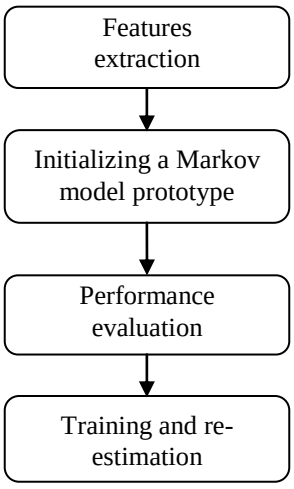


Fig 3. Experimental configuration of HMM model under HTK platform.

III. EXPERIMENTAL RESULTS

The Electrocardiogram records from the MIT /BIH database [8] were used in this study. We used only 18 healthy signals from different individuals. The duration of each recording is 30 minutes and the sampling frequency is 360 Hz. An extraction of the the Cepstral coefficients from each heart beat was made and the LFCC-RASTAMAT package [15] was used.

```
---------------------- Overall Results ----------------------
SENT: %Correct=99.33 [H=143, S=1, N=144]
WORD: %Corr=99.33, Acc=99.33 [H=143, D=0, S=1, I=0, N=144]
---------------------- Confusion Matrix ----------------------
       P   P   P   P   P   P   P   P   P   P   P   P   P   P   P   P   P
       1   2   3   4   5   6   7   8   9   1   1   1   1   1   1   1   1
                                           0   1   2   3   4   5   6   7   8   Del [ %c / %e]
  P1   8   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0     0
  P2   0   8   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0     0
  P3   0   0   8   0   0   0   0   0   0   0   0   0   0   0   0   0   0     0
  P4   0   0   0   8   0   0   0   0   0   0   0   0   0   0   0   0   0     0
  P5   0   0   0   0   8   0   0   0   0   0   0   0   0   0   0   0   0     0
  P6   0   0   0   0   0   7   0   0   0   0   0   0   0   0   0   1   0     0 [87.5/0.7]
  P7   0   0   0   0   0   0   8   0   0   0   0   0   0   0   0   0   0     0
  P8   0   0   0   0   0   0   0   8   0   0   0   0   0   0   0   0   0     0
  P9   0   0   0   0   0   0   0   0   8   0   0   0   0   0   0   0   0     0
 P10   0   0   0   0   0   0   0   0   0   8   0   0   0   0   0   0   0     0
 P11   0   0   0   0   0   0   0   0   0   0   8   0   0   0   0   0   0     0
 P12   0   0   0   0   0   0   0   0   0   0   0   8   0   0   0   0   0     0
 P13   0   0   0   0   0   0   0   0   0   0   0   0   8   0   0   0   0     0
 P14   0   0   0   0   0   0   0   0   0   0   0   0   0   8   0   0   0     0
 P15   0   0   0   0   0   0   0   0   0   0   0   0   0   0   8   0   0     0
 P16   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   8   0     0
 P17   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   8     0
 P18   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0     8
Ins    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
=================================================================
```

Fig 4. Confusion Matrix.

TABLE I. RECOGNITION RATE FOR DIFFERENT NUMBER OF COEFFICIENTS.

	Number of coefficients											
	12	*14*	*16*	*18*	*20*	*22*	*24*	*26*	*28*	*30*	*32*	*34*
Recognition rates	88.5%	89.17%	94.14%	95.50%	96.12%	96.33%	96.05%	97.25%	97.33%	98.61%	**99.31%**	98.05%

TABLE II. RECOGNITION RATE FOR DIFFERENT NUMBER OF FREQUENCY BANDS WITH 5 STATES AND 1 GAUSSIAN.

	Number of frequency bands											
	12	*14*	*16*	*18*	*20*	*22*	*24*	*26*	*28*	*30*	*32*	*34*
Recognition rates	94.4%	96.53%	97.05%	97.23%	98.17%	98.33%	**99.31%**	**99.31%**	**99.31%**	**99.31%**	**99.31%**	**99.31%**

In order to find the best results, the number of coefficients and the number of bank filter were varied from 12 to 34.

Table 1 presented the variation of human recognition rate for different number of coefficients with one gaussian and 5 states. It is to be noticed that 32 coefficients gave a better result. Then, the number of bank filter by window was varied for 32 coefficients. One can note that the rate became constant from 24 filter banks. Then it decreased in thirty the fourth bank filter.Table 2 illustrated those results.

Thus, the best configuration of cepstral features extraction used in this paper is 32 cepstral coefficients, 24 frequency bands and the width of window is 0.5 seconds with 50% overlap.

Then, the cepstral coefficients using HMM were classified. For this, an approach two-thirds one third was used. That is to say, 20 minutes of ECG records were used for training and the last 10 minutes for testing.

Assessments for improved model topology of classifier, the number of Gaussian were varied between 1 to 5 and the number of state from 5 to 7. The best rates are obtained for HMM topology with 1 Gaussian and 5 states which is the Bakis model.

Table 3 presented the recognition rates for 32 coefficients with different states and several number Gaussians. The best rate reached 99.31% for 1 Gaussian and 5 states. The more the number of Gaussians and number of states increase, the more the rate decreases.

TABLE III. RECOGNITION RATES FOR DIFFERENT STATES WITH SEVERAL GAUSSIAN NUMBERS PER STATES.

	1 Gaussian	**3 Gaussians**	**5 Gaussians**
5 States	**99.31%**	95.33%	88.67%
6 States	95.50%	89.17%	88.33%
7 States	92.83%	90%	87.50%

In order to improve the results, dynamic features which are first order delta (Δ) and energy (E) were added to the coefficients. The experimental results are illustrated in Table2 which reports the recognition rates obtained for each case.
As it can be seen, the best performance of our system has been obtained by 32 cepstral coefficients plus its Energy that make a 33 dimensional feature vector and the result is reaching 99.33%.

TABLE IV. RECOGNITION RATE FOR DIFFERENT NUMBER OF FREQUENCY BANDS WITH 5 STATES AND 1 GAUSSIAN.

LFCC	**LFCC+ E**	**LFCC+Δ**	**LFCC+E+Δ**
93.31%	**99.33%**	92.50%	91.67%

In previous work [2], the temporal parameters which are morphological descriptors and Hermite Polynomials Expansion Coefficients (HPEc) were classified with Hidden Model Markov (HMM).
According to the obtained results concerning this paper, the study concludes that the cepstral information in the homomorphic domain has given a human recognition rate that is equal 99.33% which remains slightly higher than that temporal information being 99.02%.

IV. CONCLUSION

We presented in this paper, a new personal identification method which used the cepstral information in the homomorphic domain. The RASTAMAT Package was used for extracting of the Lineair Frequency Cepstral Coefiicients.

Electrocardiogram (ECG) data for this investigation was obtained from MIT BIH database. Using this method, 99.33% human recognition rate was achieved for 18 healthy subjects. An interesting rate for identifying normal people is obtained by adding Energy to cepstral coefficients. That's why this study could be considered as a motivation towards the use of the cepstral information in the Human recognition system.

For future identification systems, the cepstral information with other parameters in time domain can be combined and multimodal biometrics can be used by associating the Electrocardiogram signal with other biometric parameters.

REFERENCES

[1] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric systems,"IEEE Transactions on 188 Circuit and Systems for Video Technology, vol. 14, no. 1, pp.4–20, 2004.

[2] E.Rabhi, and Z. Lachiri," Biometric Personal Identification System using the ECG Signal," Computing in Cardiology Conference (CinC), vol.40,pp.507–510,2013.

[3] G.D.Clifford," A Novel Framework for Signal Representation and Source Separation,"Applications to Filtering and Segmentation of Biosignals, Journal of Biological Systems, Vol. 14, No. 2, pp. 169-183 ,June 2006.

[4] L. Biel, O. Pettersson, L. Philipson, and P. Wide, "ECG analysis: a new approach in human identification," IEEE Transactions on Instrumentation and Measurement, vol. 50,no. 3, pp. 808–812, 2001.

[5] L.R. Rabiner," A tutorial on hidden markov models and selected applications in speech recognition," Proceedings of the IEEE,vol.7,pp.257–286,1989.

[6] M.Li, S.Narayanan," Robust ECG Biometrics by Fusing Temporal and Cepstral Information," Pattern Recognition (ICPR), pp.1326-1329, 2010.

[7] M.Li, V. Rozgic, G.Thatte, S.Lee, A.Emken, M.Annavaram, U. Mitra, D.Spruijt-Metz and S.Narayanan, "Multimodal Physical Activity Recognition by Fusing Temporal and Cepstral Information," IEEE Transactions on Neural Systems & Rehabilitation Engineering, vol 18, issue4, August, 2010.

[8] MIT-BIH Normal Sinus Rhythm Database, available online: http://www.physionet.org/physiobank/database.

[9] N. Belgacem, F.B. REGUIG and A.NAIT-ALI, "Person Identification System Based on Electrocardiogram Signal Using LabVIEW, " International Journal on Computer Science and Engineering (IJCSE), Vol. 4 No. 06 June 2012.

[10] S. A. Israel, J. M. Irvine, A. Cheng, M. D. Wiederhold, and B. K. Wiederhold, "ECG to identify individuals," Pattern Recognition, vol. 38, no. 1, pp. 133–142, 2005.

[11] S.Young," The HTK Hidden Markov Model Toolkit," Design and Philosophy. Cambridge University Engineering Department - Technical report 152, 1994.

[12] S.Young, G. .Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book 3.2, Cambridge University Engineering Department - Speech Group, 2001-2006.

[13] T. W. Shen, W. J. Tompkins, and Y. H. Hu, "One-lead ECG for identity verifica-tion," in Proc. of the 2nd Conf. of the IEEE Eng. in Med. and Bio. Society and the Biomed. Eng. Society, vol. 1, pp. 62–63,2002.

[14] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, S. Shamma, "Linear versus Mel- Frequency cepstral coefficients for speaker recognition", ASRU 2011 (IEEE Automatic Speech Recognition and Understanding Workshop).

[15] X. Zhou (2011), LFCC- RASTA in Matlab, available online: http://terpconnect.umd.edu/~zxinhui/

Incorporating Auxiliary Information in Collaborative Filtering Data Update with Privacy Preservation

Xiwei Wang, Jun Zhang,
Pengpeng Lin, Nirmal Thapa
Dept. of Computer Science
University of Kentucky
Lexington, Kentucky 40506-0633
Email: {xiwei, jzhang}@cs.uky.edu,
{m.lin, nirmalthapa}@uky.edu

Yin Wang
Dept. of Math and Computer Science
Lawrence Technological University
Southfield, Michigan, 48075
Email: ywang12@ltu.edu

Jie Wang
Dept. of Computer Information Systems
Indiana University Northwest
Gary, Indiana, 46408
Email: wangjie@iun.edu

***Abstract*—Online shopping has become increasingly popular in recent years. More and more people are willing to buy products through Internet instead of physical stores. For promotional purposes, almost all online merchants provide product recommendations to their returning customers. Some of them ask professional recommendation service providers to help develop and maintain recommender systems while others need to share their data with similar shops for better product recommendations. There are two issues, (1) how to protect customers' privacy while retaining data utility before they release the data to the third parties; (2) based on (1), how to handle data growth efficiently.**

In this paper, we propose a NMF (Nonnegative Matrix Factorization)-based data update approach in collaborative filtering (CF) that solves the problems. The proposed approach utilizes the intrinsic property of NMF to distort the data for protecting user's privacy. In addition, the user and item auxiliary information is taken into account in incremental nonnegative matrix tri-factorization to help improve the data utility. Experiments on three different datasets (MovieLens, Sushi and LibimSeTi) are conducted to examine the proposed approach. The results show that our approach can quickly update the new data and provide both high level privacy protection and good data utility.

***Keywords*—auxiliary information; collaborative filtering; data growth; nonnegative matrix factorization; privacy**

I. Introduction

The emergence of E-commerce not only helps sellers save resources and time but also facilitates online transactions. Different kinds of promotions have been adopted by merchants to advertise their products. Conventional stores usually present popular products, e.g., batteries, gift cards, and magazines at the checkout line besides offering discounts, which is a typical way of product recommendations. For returning customers, online stores are far superior with respect to product recommendation since they use users'[1] purchase history in recommender system to achieve accurate recommendation. The so called recommender system is a program that utilizes algorithms to predict users' purchase interests by profiling their shopping patterns. Most popular recommender systems utilize CF techniques, e.g., item/user correlation based CF [22], SVD (Singular Value Decomposition) based latent factor CF [24], and NMF (Nonnegative Matrix Factorization) based CF [34], [4].

In many online recommender systems, it is inevitable for data owners to expose their data to other parties. For instance, due to the lack of easy-to-use technology, some online merchants buy services from professional recommendation service providers to help build their recommender systems. In addition, many shops share their real time data with partners for better product recommendations. Such examples include two or more online book stores that sell similar books, and online movie rental websites that have similar movies in their systems. In these scenarios, exposed data can cause privacy leakage of user information if no preprocessing is done. Typical privacy information includes the ratings of a user left on particular items and on which items that this user has rated. People would not like others (except the website where they purchased the products because they have no choice) to know what they are interested in and to what extent they like or dislike the items. This is the most fundamental privacy problem in collaborative filtering. Thus privacy preserving collaborative filtering algorithms [3], [21], [19] were proposed to tackle the problem.

Most CF algorithms work on user-item rating matrices to make recommendations. These numerical matrices store user's ratings on particular items, typically with users corresponding to the rows and items corresponding to the columns. In general, the rating matrices are very sparse, meaning that there are lots of missing values. Therefore, two tasks need to be done before a data owner (merchant) releases the data to a third party: missing value imputation and data perturbation[2].

Furthermore, data owners are responsible for efficiently handling the fast growth of data. Once new data arrives, data owners need to perform incremental data update and send the imputed and perturbed data to the third parties. To this end, Wang and Zhang[30] proposed an SVD-based privacy preserving data update scheme to handle data growth efficiently and preserve privacy as well. Nevertheless, their SVD-based update scheme has a few deficiencies: (1) The SVD algorithm

[1]The terms "customer" and "user" will be used interchangeably as they refer to the same thing in this context. Same convention applies to "product" and "item".

[2]Data perturbation is a form of privacy-preserving data mining technique. It falsifies the data before publication by introducing error to elements purposely for confidentiality reasons [8]. Data perturbation is widely used in collaborative filtering for privacy preservation.

cannot be applied to incomplete matrix so missing values imputation is required. Choosing a good imputation method is not quite straightforward and it is dependant on different datasets. (2) The update scheme only utilizes the rating data while ignores other auxiliary information. It is known that in some datasets, e.g., MovieLens dataset [24], Sushi preference dataset [12] and LibimSeTi Dating Agency (LibimSeTi for short) dataset [2], auxiliary information of users or items, e.g., user's demographic data, item's categorical data, is also provided. This information, if properly used, can improve the recommendation accuracy especially when the original rating matrix is extremely sparse. (3) The time complexity of their method contains a cubic term with respect to the number of new rows or columns. It is a potentially expensive factor in the update process, especially when a large amount of new data comes in.

In this paper, we propose a NMF-based data update approach that solves the issues. The approach, named iAux-NMF is based on the incremental nonnegative matrix tri-factorization algorithms [7]. We start with computing the weighted and constrained nonnegative matrix tri-factorization for the original sparse rating matrix (with a lot of missing values), utilizing both the rating matrix itself and the auxiliary information. The factor matrices of NMF are then used to approximate the original rating matrix with missing values imputed. Meanwhile, the data is automatically perturbed due to the intrinsic properties of NMF [29]. For new data, iAux-NMF is performed to produce imputed and perturbed data. This process can conceal which items the users have rated as there is no more missing entries and disguise the true rating values since the processed ratings and the original ones are different. By doing so, even though the third party has this data in its hand, it does not know which ratings it can trust or to what extent it can trust. Therefore, user's privacy is protected.

We examine our approach in several aspects: (1) correctness of the approximated rating matrix, (2) clustering analysis on the approximated rating matrix for investigating user rating distribution, (3) privacy level of the approximated rating matrix, (4) time cost of the algorithms, and (5) parameter study. The results demonstrate that our approach imputes and perturbs the new data in a timely manner with satisfying privacy level and high data utility (less compromised data accuracy). The processed data is also reasonable from the clustering point of view.

The contributions of this paper are threefold:

1) No particular missing value imputation methods required during the data update;
2) Incorporating auxiliary information into the update process to improve data utility;
3) Higher data update efficiency.

The remainder of this paper is organized as follows. Section II gives the related work. Section III defines the problem and related notations. Section IV describes the main idea of the proposed approach. Section V presents the experiments and discusses the results. Some concluding remarks and future work are given in VI.

II. Related Work

Privacy preserving data update was first studied by Wang et al.[28] who presented a data value hiding method for clustering algorithms based on incremental SVD technique [26]. Their method can produce a significant increase in speed for the SVD-based data value hiding model, better scalability, and better real-time performance of the model. Motivated by their work, Wang and Zhang[30] incorporated the missing value imputation and randomization-based perturbation as well as a post-processing procedure into the incremental SVD to update the new data with privacy preservation in collaborative filtering.

Besides SVD, NMF has also been studied in collaborative filtering. Zhang et al.[34] applied NMF to collaborative filtering to learn the missing values in the rating matrix. They compared an expectation maximization (EM) based procedure (using NMF as its solution) with the weighted nonnegative matrix factorization (WNMF) based method which was previously applied to missing value imputation in matrix of network distances [18]. By integrating the advantages of both algorithms, they presented a hybrid method and demonstrated its effectiveness on real datasets. Chen et al.[4] proposed an orthogonal nonnegative matrix tri-factorization (ONMTF) [7] based collaborative filtering algorithm. Their algorithm also took into account the user similarity and item similarity. Our approach is generally based on the nonnegative matrix tri-factorization (NMTF) but we add further constraints to the objective function.

NMF with additional constraints has been applied to different fields. Li et al.[16] proposed nonnegative matrix factorization with orthogonality constraints for detection of a target spectrum in a given set of Raman spectra data. Hoyer et al.[10] extended NMF by adding a sparsity-inducing penalty to the objective function to include the option for explicit sparseness control. Ferdowsi et al.[9] proposed a constrained NMF algorithm for separation of active area in the brain from fMRI. In their work, prior knowledge of the sensory stimulus is incorporated into standard NMF to find new update rules for the decomposition process.

Thapa et al.[25] proposed explicit incorporation of the additional constraint, called "clustering constraint", into NMF in order to suppress the data patterns in the process of performing the matrix factorization. Their work is based on the idea that one of the factor matrices in NMF contains cluster membership indicators. The clustering constraint is another indicator matrix with altered class membership in it. This constraint then guides NMF in updating factor matrices. Enlightened by that paper, we convert users' and items' auxiliary information into cluster membership indicator matrices and apply them to NMTF as additional constraints. We do not hide data pattern, but update factor matrices in a more reasonable way for better missing value imputation.

III. Problem Description

Assume the data owner has three matrices: a sparse user-item rating matrix (denoted by $R \in \mathbb{R}^{m \times n}$), a user feature matrix (denoted by $F_U \in \mathbb{R}^{m \times k_U}$), and an item feature matrix (denoted by $F_I \in \mathbb{R}^{n \times k_I}$), where there are m users, n items, k_U user features, and k_I item features. An entry r_{ij} in R

represents the rating left on item j by user i. The valid range of rating value varies from website to website. Some use the $1 \sim 5$ scale with 1 as the lowest rating (most disliked) and 5 as the highest rating (most favorated) while some others use the $-10 \sim 10$ scale with -10 as the lowest rating, 0 as neutral rating, and 10 as the highest rating.

The original rating matrix contains the real rating values left by users on items, which means it can be used to identify the shopping patterns of users. These patterns can reveal some user privacy, so releasing the original rating data without any privacy protection will cause the privacy breach. One possible way to protect the user privacy before releasing the rating matrix is to impute the matrix and then perturb it. In this procedure, imputation estimates the missing ratings as well as conceals the user preference on particular items (no missing value means there is no way to tell which items have been rated by users since all items are marked as rated.) while the perturbation distorts the ratings so that user's preferences on particular items are blurred.

As for the user feature matrix F_U and item feature matrix F_I, they contain users' and items' information, respectively. They are taken into account to help impute the missing entries in rating matrix for better accuracy. The processed (imputed and perturbed) matrix, denoted by $R_r \in \mathbb{R}^{m \times n}$ is the one that will be handed over to the third party.

When new users' transactions arrive, the new rows (each row contains the ratings left on items by the corresponding user), denoted by $T \in \mathbb{R}^{p \times n}$, should be appended to the original matrix R. Meanwhile, this new users' auxiliary information is also available, and thus the feature matrix is updated as well, i.e.,

$$\begin{bmatrix} R \\ T \end{bmatrix} \rightarrow R' \qquad \begin{bmatrix} F_U \\ \Delta F_U \end{bmatrix} \rightarrow F'_U \tag{1}$$

where $\Delta F_U \in \mathbb{R}^{p \times k_U}$.

Similarly, when new items arrive, the new columns (each column contains the ratings left by users on the corresponding item), denoted by $G \in \mathbb{R}^{m \times q}$, should be appended to the original matrix R, so should the item feature matrix, i.e.,

$$[\ R \quad G\ ] \rightarrow R'', \qquad \begin{bmatrix} F_I \\ \Delta F_I \end{bmatrix} \rightarrow F'_I \tag{2}$$

where $\Delta F_I \in \mathbb{R}^{q \times k_I}$.

To protect users' privacy, the new rating data must be processed before it is released. We use $T_r \in \mathbb{R}^{p \times n}$ to denote the processed new rows and $G_r \in \mathbb{R}^{m \times q}$ for processed new columns.

IV. Using iAux-NMF for Privacy Preserving Data Update

In this section, we will introduce the iAux-NMF (incremental auxiliary nonnegative matrix factorization) algorithm and its application in incremental data update with privacy preservation.

A. Aux-NMF

While iAux-NMF deals with the incremental data update, we want to present the non-incremental version, named Aux-NMF beforehand. This section is organized as follows: developing the objective function, deriving the update formula, convergence analysis, and the detailed algorithm.

1) Objective Function: Nonnegative matrix factorization (NMF)[15] is a widely used dimension reduction method in many applications such as clustering [7], [13], text mining [31], [20], image processing and analysis [33], [23], data distortion based privacy preservation [11], [25], etc. NMF is also applied in collaborative filtering to make product recommendations [34], [4].

A conventional NMF is defined as follows [15],

$$R_{m \times n} \approx U_{m \times k} \cdot V^T_{n \times k} \tag{3}$$

The goal is to find a pair of orthogonal nonnegative matrices U and V (i.e., $U^T U = I, V^T V = I$) that minimize the Frobenius norm (or Euclidean norm) $\|R - UV^T\|_F$. It comes up with the objective function

$$min_{U \geq 0, V \geq 0} f(R, U, V) = \|R - UV^T\|^2_F \tag{4}$$

In this paper, we want to develop a NMF-based matrix factorization technique which takes into account the weight and constraint. It is expected to preserve the data privacy by imputing and perturbing the values during its update process.

It is worth noting that one of the significant distinctions between collaborative filtering data and other data is the missing value issue. One user may have rated only a few items and one item may receive only a small number of ratings. It results in a very sparse rating matrix which cannot be simply fed to the matrix factorization algorithms, such as SVD and NMF. Those missing values should be imputed properly during the pre-processing step. Existing imputation methods include random value imputation, mean value imputation [24], EM (Expectation Maximization) imputation [5], [32], and linear regression imputation [27], etc. Nevertheless, all of them require extra time to compute the missing values. In contrast, weighted NMF (WNMF) [34] can work with sparse matrix without separate imputation.

Given a weight matrix $W \in \mathbb{R}^{m \times n}$ that indicates the value existence in the rating matrix R (see Eq. (6)), the objective function of WNMF is

$$min_{U \geq 0, V \geq 0} f(R, W, U, V) = \|W \circ (R - UV^T)\|^2_F \tag{5}$$

where $\circ$ denotes the element-wise multiplication.

$$w_{ij} = \begin{cases} 1 & if \quad r_{ij} \neq 0 \\ 0 & if \quad r_{ij} = 0 \end{cases} \quad (w_{ij} \in W, r_{ij} \in R) \tag{6}$$

When WNMF converges, $\tilde{R} = UV^T$ is the matrix with all missing entries filled. Since the residual exists, $\tilde{R}$ is different from R, making it a perturbed version of R. As we stated in Section I, users do not want their privacy, i.e., their ratings left on particular items and on which items they have rated, to be released to other people. In WNMF, both of them are protected.

In [6], Ding et al. showed the equivalency between NMF and K-Means clustering. When given a matrix R with objects as rows and attributes as columns, the two matrices U and V produced by NMF on R describe the clustering information of the objects: each column vector of U, u_i, can be regarded as a basis and each data point r_i is approximated by a linear combination of these k bases, weighted by the components of V [17], where k is the rank of factor matrices. Thus the objects are grouped into clusters in terms of matrix U.

However, in some cases, the data matrix R can represent relationships between two sorts of objects, e.g., a user-item rating matrix in collaborating filtering applications and a term-document matrix in text mining applications. It is expected that both row (user/term) clusters and column (item/document) clusters can be obtained by performing NMF on R. Due to the intrinsic property of NMF, it is very difficult to find two matrices U and V that represent user clusters and item clusters respectively at the same time. Hence, an extra factor matrix is needed to absorb the different scales of R, U, V for simultaneous row clustering and column clustering [7]. Eq. (7) gives the objective function of NMTF(Nonnegative Matrix Tri-Factorization).

$$min_{U\geq 0,S\geq 0,V\geq 0} f(R,U,S,V) = \|R - USV^T\|_F^2 \tag{7}$$

where $U \in \mathbb{R}^{m\times k}$, $S \in \mathbb{R}^{k\times l}$, and $V \in \mathbb{R}^{n\times l}$ (U and V are orthogonal matrices).

The use of S brings in a large scale of freedom for U and V so that they can focus on row and column clustering and preserve more privacy during the factorization process. In this scheme, both U and V are cluster membership indicator matrices while S plays the role of coefficient matrix. Note that objects corresponding to rows in R are clustered into k groups and objects corresponding to columns are clustered into l groups.

With auxiliary information of users and items, we can convert the NMTF to a supervised learning process by applying cluster constraints to the objective function (7), i.e.,

$$\begin{aligned} &min_{U\geq 0,S\geq 0,V\geq 0} f(R,U,S,V,C_U,C_I) = \\ &\alpha\cdot\|R - USV^T\|_F^2 + \beta\cdot\|U - C_U\|_F^2 \\ &+\gamma\cdot\|V - C_I\|_F^2 \end{aligned} \tag{8}$$

where α, β, and γ are coefficients that control the weight of each part. C_U and C_I are user cluster matrix and item cluster matrix. They are obtained by running K-Means clustering algorithm on user feature matrix F_U and item feature matrix F_I as mentioned in Section III.

Combining (5) and (8), we develop the objective function for weighted and constrained nonnegative matrix tri-factorization, as

$$\begin{aligned} &min_{U\geq 0,S\geq 0,V\geq 0} f(R,W,U,S,V,C_U,C_I) = \\ &\alpha\cdot\|W\circ(R - USV^T)\|_F^2 + \beta\cdot\|U - C_U\|_F^2 \\ &+\gamma\cdot\|V - C_I\|_F^2. \end{aligned} \tag{9}$$

We name this matrix factorization Aux-NMF, indicating that it incorporates the user/item auxiliary information into the factorization.

2) Update Formula: In this section, we illustrate the derivation of update formulae for Aux-NMF.

Let $L = f(R,W,U,S,V,C_U,C_I)$, $X = \|W\circ(R - USV^T)\|_F^2$, $Y = \|U - C_U\|_F^2$, and $Z = \|V - C_I\|_F^2$. Take derivative of X with respect to U, S, and V:

$$\frac{\partial X}{\partial U} = -2(W\circ R)VS^T + 2W\circ(USV^T)VS^T \tag{10}$$

$$\frac{\partial X}{\partial S} = -2U^T(W\circ R)V + 2U^T[W\circ(USV^T)]V \tag{11}$$

$$\frac{\partial X}{\partial V} = -2(W\circ R)^TUS + 2[W\circ(USV^T)]^TUS \tag{12}$$

Take derivative of Y with respect to U, S, and V:

$$\frac{\partial Y}{\partial U} = 2U - 2C_U, \quad \frac{\partial Y}{\partial S} = \frac{\partial Y}{\partial V} = 0 \tag{13}$$

Take derivative of Z with respect to U, S, and V:

$$\frac{\partial Z}{\partial U} = \frac{\partial Z}{\partial S} = 0, \quad \frac{\partial Z}{\partial V} = 2V - 2C_I \tag{14}$$

Using (10) to (14), we get the derivatives of L:

$$\begin{aligned} \frac{\partial L}{\partial U} &= 2\alpha[W\circ(USV^T)]VS^T + 2\beta U \\ &\quad - 2\alpha(W\circ R)VS^T - 2\beta C_U \end{aligned} \tag{15}$$

$$\begin{aligned} \frac{\partial L}{\partial V} &= 2\alpha[W\circ(USV^T)]^TUS + 2\gamma V \\ &\quad - 2\alpha(W\circ R)^TUS - 2\gamma C_I \end{aligned} \tag{16}$$

$$\begin{aligned} \frac{\partial L}{\partial S} &= 2\alpha U^T[W\circ(USV^T)]V \\ &\quad - 2\alpha U^T(W\circ R)V \end{aligned} \tag{17}$$

To obtain update formula, we use the Karush-Kuhn-Tucker (KKT) complementary condition [14] for the nonnegativity of U, S, and V. We have

$$\begin{aligned} &\{2\alpha[W\circ(USV^T)]VS^T + 2\beta U \\ &\quad - 2\alpha(W\circ R)VS^T - 2\beta C_U\}_{ij}U_{ij} = 0 \end{aligned} \tag{18}$$

$$\begin{aligned} &\{2\alpha[W\circ(USV^T)]^TUS + 2\gamma V \\ &\quad - 2\alpha(W\circ R)^TUS - 2\gamma C_I\}_{ij}V_{ij} = 0 \end{aligned} \tag{19}$$

$$\{2\alpha U^T[W\circ(USV^T)]V - 2\alpha U^T(W\circ R)V\}_{ij}S_{ij} = 0 \tag{20}$$

They give rise to the corresponding update formulae:

$$U_{ij} = U_{ij}\cdot\frac{[\alpha(W\circ R)VS^T + \beta C_U]_{ij}}{\{\alpha[W\circ(USV^T)]VS^T + \beta U\}_{ij}} \tag{21}$$

$$V_{ij} = V_{ij}\cdot\frac{[\alpha(W\circ R)^TUS + \gamma C_I]_{ij}}{\{\alpha[W\circ(USV^T)]^TUS + \gamma V\}_{ij}} \tag{22}$$

$$S_{ij} = S_{ij}\cdot\frac{[U^T(W\circ R)V]_{ij}}{\{U^T[W\circ(USV^T)]V\}_{ij}} \tag{23}$$

Assume $k, l \ll \min(m,n)$, the time complexities of updating U, V, and S in each iteration are all $O(mn(k+l))$. Therefore, the time complexity of Aux-NMF in each iteration is $O(mn(k+l))$.

3) Convergence Analysis: We follow [15] to prove that the objective function L is nonincreasing under the update formulas (21), (22), and (23).

Definition 1: $H(u, u')$ is an auxiliary function for $F(u)$ if the conditions

$$H(u, u') \geq F(u), \quad H(u, u) = F(u) \tag{24}$$

are satisfied.

Lemma 1: If H is an auxiliary function for F, then F is nonincreasing under the update

$$u^{t+1} = \underset{u}{\operatorname{argmin}} H(u, u^t) \tag{25}$$

Lemma 1 can be easily proved since we have $F(u^{t+1}) = H(u^{t+1}, u^{t+1}) \leq H(u^{t+1}, u^t) \leq H(u^t, u^t) = F(u^t)$.

We will prove the convergences of the update formulas (21), (22), and (23) by showing that they are equivalent to (25), with proper auxiliary functions defined.

Let us rewrite the objective function L,

$$\begin{aligned} L = & tr[\alpha(W \circ R)^T \cdot (W \circ R)] \\ & + tr\{-2\alpha(W \circ R)^T \cdot [W \circ (USV^T)]\} \\ & + tr\{\alpha[W \circ (USV^T)]^T \cdot [W \circ (USV^T)]\} \\ & + tr(\beta U^T U) + tr(-2\beta U^T C_U) + tr(\beta C_U^T C_U) \\ & + tr(\gamma V^T V) + tr(-2\gamma V^T C_I) + tr(\gamma C_I^T C_I) \end{aligned} \tag{26}$$

where $tr(*)$ is the trace of a matrix.

Eliminating the irrelevant terms, we define the following functions that are only related to U, V, and S, respectively.

$$\begin{aligned} L(U) = & tr\{-2\alpha(W \circ R)^T \cdot [W \circ (USV^T)] \\ & + \alpha[W \circ (USV^T)]^T \cdot [W \circ (USV^T)] \\ & + \beta U^T U - 2\beta U^T C_U\} \\ = & tr\{[-2[\alpha(W \circ R)VS^T + \beta C_U]U^T \\ & + U^T[\alpha W \circ (USV^T)VS^T] + U^T(\beta U)\} \end{aligned} \tag{27}$$

$$\begin{aligned} L(V) = & tr\{-2\alpha(W \circ R)^T \cdot [W \circ (USV^T)] \\ & + \alpha[W \circ (USV^T)]^T \cdot [W \circ (USV^T)] \\ & + \gamma V^T V - 2\gamma V^T C_I \\ = & tr\{[-2[\alpha(W \circ R)^T US + \gamma C_I]V^T \\ & + V^T[\alpha(W \circ (USV^T))^T US] + V^T(\gamma V)\} \end{aligned} \tag{28}$$

$$\begin{aligned} L(S) = & tr\{-2\alpha(W \circ R)^T \cdot [W \circ (USV^T)] \\ & + \alpha[W \circ (USV^T)]^T \cdot [W \circ (USV^T)] \\ = & tr\{[-2\alpha U^T(W \circ R)V]S^T \\ & + [\alpha U^T(W \circ (USV^T))V]S^T\} \end{aligned} \tag{29}$$

Lemma 2: For any matrices $X \in \mathbb{R}_+^{n \times n}$, $Y \in \mathbb{R}_+^{k \times k}$, $F \in \mathbb{R}_+^{n \times k}$, $F' \in \mathbb{R}_+^{n \times k}$, and X, Y are symmetric, the following inequality holds

$$\sum_{i=1}^{n}\sum_{j=1}^{k} \frac{(XF'Y)_{ij}F_{ij}^2}{F'_{ij}} \geq tr(F^T XFY) \tag{30}$$

The proof of Lemma 2 is presented in [7]. We will use this lemma to build an auxiliary function for $L(U)$ (since it is similar to $L(V)$ and $L(S)$, we will not discuss the convergences for them).

Lemma 3:

$$\begin{aligned} H(U, U') = & -2\sum_{ij}\{[\alpha(W \circ R)VS^T + \beta C_U]U^T\}_{ij} \\ & + \sum_{ij} \frac{(\alpha W \circ (U'SV^T)VS^T + \beta U')_{ij}U_{ij}^2}{U'_{ij}} \end{aligned} \tag{31}$$

is an auxiliary function of $L(U)$ and the global minimum of $H(U, U')$ can be achieved by

$$U_{ij} = U'_{ij} \cdot \frac{[\alpha(W \circ R)VS^T + \beta C_U]_{ij}}{\{\alpha[W \circ (U'SV^T)]VS^T + \beta U'\}_{ij}} \tag{32}$$

Proof: We need to prove two conditions as specified in Definition 1. It is apparent that $H(U, U) = L(U)$. According to Lemma 2, we have

$$\begin{aligned} & \sum_{ij} \frac{(\alpha W \circ (U'SV^T)VS^T + \beta U')_{ij}U_{ij}^2}{U'_{ij}} \\ & = \sum_{ij} \frac{(\alpha W \circ (U'SV^T)VS^T)_{ij}U_{ij}^2}{U'_{ij}} + \sum_{ij} \frac{(\beta U')_{ij}U_{ij}^2}{U'_{ij}} \\ & \geq tr\{U^T[\alpha W \circ (USV^T)VS^T]\} + tr[U^T(\beta U)] \end{aligned} \tag{33}$$

I.e., $H(U, U') \geq L(U)$. Thus $H(U, U')$ is an auxiliary function of $L(U)$.

To find the global minimum of $H(U, U')$ with U' fixed, we take derivative of $H(U, U')$ with respect to U_{ij} and let it be zero:

$$\begin{aligned} \frac{\partial H(U, U')}{\partial U_{ij}} = & \{-2[\alpha(W \circ R)VS^T + \beta C_U]\}_{ij} \\ & + 2\frac{(\alpha W \circ (U'SV^T)VS^T + \beta U')_{ij}U_{ij}}{U'_{ij}} = 0 \end{aligned} \tag{34}$$

Solving for U_{ij}, we have

$$U_{ij} = U'_{ij} \cdot \frac{[\alpha(W \circ R)VS^T + \beta C_U]_{ij}}{\{\alpha[W \circ (U'SV^T)]VS^T + \beta U'\}_{ij}} \tag{35}$$

Since the Hessian matrix $\partial^2 H(U, U')/\partial U_{ij}\partial U_{kl}$ is positive definite, $H(U, U')$ is a convex function and the minimum obtained by Eq. (35) is also the global minimum. ■

Similarly, the convergences of update formulas (23) and (22) can be proved as well.

4) Detailed Algorithm: In this section, we present the specific algorithm for Aux-NMF in collaborating filtering which is the basis of incremental Aux-NMF.

Algorithm 1 depicts the whole process of performing Aux-NMF on a rating matrix.

Though Aux-NMF will eventually converge to a local minimum, it may take hundreds or even thousands of iterations. In our algorithm, we set an extra stop criterion - the maximum iteration counts. In collaborative filtering, this value varies from $10 \sim 100$ and can generally produce good results.

Algorithm 1 Aux-NMF

Require:
User-Item rating matrix: $R \in \mathbb{R}^{m \times n}$;
User feature matrix: $F_U \in \mathbb{R}^{m \times k_U}$;
Item feature matrix: $F_I \in \mathbb{R}^{n \times k_I}$;
Column dimension of U: k;
Column dimension of V: l;
Coefficients in objective function: α, β, and γ;
Number of maximum iterations: $MaxIter$.
Ensure:
Factor matrices: $U \in \mathbb{R}^{m \times k}$, $S \in \mathbb{R}^{k \times l}$, $V \in \mathbb{R}^{n \times l}$;
User cluster membership indicator matrix: $C_U \in \mathbb{R}^{m \times k}$;
Item cluster membership indicator matrix: $C_I \in \mathbb{R}^{n \times l}$;
User cluster centroids: $Centroids_U$;
Item cluster centroids: $Centroids_I$;

1: Cluster users into k groups based on F_U by K-Means algorithm $\rightarrow C_U$, $Centroids_U$;
2: Cluster items into l groups based on F_I by K-Means algorithm $\rightarrow C_I$, $Centroids_I$;
3: Initialize U, S, and V with random values;
4: Build weight matrix W by Eq. (6);
5: Set $iteration = 1$ and $stop = false$;
6: **while** $(iteration < MaxIter)$ *and* $(stop == false)$ **do**
7: $U_{ij} \leftarrow U_{ij} \cdot \frac{[\alpha(W \circ R)VS^T + \beta C_U]_{ij}}{\{\alpha[W \circ (USV^T)]VS^T + \beta U\}_{ij}}$;
8: $V_{ij} \leftarrow V_{ij} \cdot \frac{[\alpha(W \circ R)^T US + \gamma C_I]_{ij}}{\{\alpha[W \circ (USV^T)]^T US + \gamma V\}_{ij}}$;
9: $S_{ij} \leftarrow S_{ij} \cdot \frac{[U^T(W \circ R)V]_{ij}}{\{U^T[W \circ (USV^T)]V\}_{ij}}$;
10: $L \leftarrow \alpha \cdot \|W \circ (R - USV^T)\|_F^2 + \beta \cdot \|U - C_U\|_F^2 + \gamma \cdot \|V - C_I\|_F^2$;
11: **if** (L increases in this iteration) **then**
12: $stop = true$;
13: Restore U, S, and V to their values in last iteration.
14: **end if**
15: **end while**
16: Return $U, S, V, C_U, C_I, Centroids_U$, and $Centroids_I$.

B. iAux-NMF

As discussed in Section III, new data can be regarded as new rows or new columns in the matrix. They are imputed and perturbed by iAux-NMF (incremental Aux-NMF) with the aid of $U, S, V, C_U, C_I, Centroids_U$, and $Centroids_I$ generated by Algorithm 1.

iAux-NMF is technically the same as Aux-NMF, but focuses on a series of new rows or new columns. Hence, in this section we will describe the incremental case of Aux-NMF by row update and column update separately.

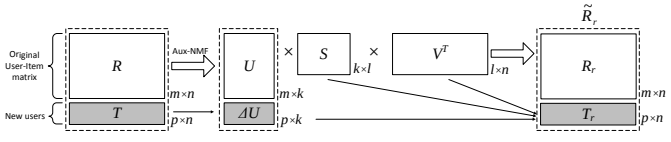


Fig. 1: Updating New Rows in iAux-NMF

1) Row/User Update: In Eq. (1), we see that $T \in \mathbb{R}^{p \times n}$ is added to R as a few rows. This process is illustrated in Fig. 1. T should be imputed and perturbed before being released. As we did in Section IV-A1, the objective function is developed here, i.e.,

$$min_{\Delta U \geq 0} f(T, W_T, \Delta U, S, V, \Delta C_U) = \alpha \cdot \|W_T \circ (T - \Delta USV^T)\|_F^2 + \beta \cdot \|\Delta U - \Delta C_U\|_F^2 \quad (36)$$

As in Section IV-A2, we obtain the update formula for this objective function, as

$$\Delta U_{ij} = \Delta U_{ij} \cdot \frac{[\alpha(W_T \circ T)VS^T + \beta \Delta C_U]_{ij}}{\{\alpha[W_T \circ (\Delta USV^T)]VS^T + \beta \Delta U\}_{ij}} \quad (37)$$

Convergence of (37) can be proved similarly as in Section IV-A3. Since row update only works on new rows, the time complexity of the algorithm in each iteration is $O(pn(l+k) + pkl)$. Assume $k, l \ll \min(p, n)$, the time complexity is then simplified to $O(pn(l+k))$.

Algorithm 2 illustrates the row update in iAux-NMF.

Algorithm 2 iAux-NMF for Row Update

Require:
New rating data: $T \in \mathbb{R}^{p \times n}$;
New user feature matrix: $\Delta F_U \in \mathbb{R}^{p \times k_U}$;
Coefficients in objective function: α, β, and γ;
Factor matrices: $U \in \mathbb{R}^{m \times k}$, $S \in \mathbb{R}^{k \times l}$, $V \in \mathbb{R}^{n \times l}$;
User cluster membership indicator matrix: $C_U \in \mathbb{R}^{m \times k}$;
User cluster centroids: $Centroids_U$;
Number of maximum iterations: $MaxIter$.
Ensure:
Updated factor matrix: $U' \in \mathbb{R}^{(m+p) \times k}$;
Updated user cluster membership indicator matrix: $C'_U \in \mathbb{R}^{(m+p) \times k}$;
Updated user cluster centroids: $Centroids'_U$;
Imputed and perturbed new data: $T_r \in \mathbb{R}^{p \times n}$;

1: Cluster new users into k groups based on ΔF_U and $Centroids_U$ by K-Means algorithm $\rightarrow$ ΔC_U, $Centroids'_U$;
2: Initialize $\Delta U \in \mathbb{R}^{p \times k}$ with random values;
3: Build weight matrix W_T by Eq. (6);
4: Set $iteration = 1$ and $stop = false$;
5: **while** $(iteration < MaxIter)$ *and* $(stop == false)$ **do**
6: $\Delta U_{ij} \leftarrow \Delta U_{ij} \cdot \frac{[\alpha(W_T \circ T)VS^T + \beta \Delta C_U]_{ij}}{\{\alpha[W_T \circ (\Delta USV^T)]VS^T + \beta \Delta U\}_{ij}}$
7: $L \leftarrow \alpha \cdot \|W_T \circ (T - \Delta USV^T)\|_F^2 + \beta \cdot \|\Delta U - \Delta C_U\|_F^2$;
8: **if** (L increases in this iteration) **then**
9: $stop = true$;
10: Restore U' to its value in last iteration.
11: **end if**
12: **end while**
13: Append ΔC_U to $C_U \rightarrow C'_U$;
14: Append ΔU to $U \rightarrow U'$;
15: Calculate $\Delta USV^T \rightarrow T_r$;
16: Return $U', C'_U, Centroids'_U$, and T_r.

2) Column/Item Update: Column update is almost identical to row update. When new data $G \in \mathbb{R}^{m \times q}$ arrives, they are updated by Algorithm 3. The time complexity for column update is $O(qm(l+k))$.

Algorithm 3 iAux-NMF for Column Update

Require:
New rating data: $G \in \mathbb{R}^{m \times q}$;
New item feature matrix: $\Delta F_I \in \mathbb{R}^{q \times k_I}$;
Coefficients in objective function: α, β, and γ;
Factor matrices: $U \in \mathbb{R}^{m \times k}$, $S \in \mathbb{R}^{k \times l}$, $V \in \mathbb{R}^{n \times l}$;
Item cluster indicator membership matrix: $C_I \in \mathbb{R}^{n \times l}$;
Item cluster centroids: $Centroids_I$;
Number of maximum iterations: $MaxIter$.
Ensure:
Updated factor matrix: $V' \in \mathbb{R}^{(n+q) \times l}$;
Updated item cluster membership indicator matrix: $C'_I \in \mathbb{R}^{(n+q) \times l}$;
Updated item cluster centroids: $Centroids'_I$;
Imputed and perturbed new data: $G_r \in \mathbb{R}^{m \times q}$;

1: Cluster new items into l groups based on ΔF_I and $Centroids_I$ by K-Means algorithm $\rightarrow \Delta C_I, Centroids'_I$;
2: Initialize $\Delta V \in \mathbb{R}^{q \times l}$ with random values;
3: Build weight matrix W_G by Eq. (6);
4: Set $iteration = 1$ and $stop = false$;
5: **while** $(iteration < MaxIter)$ *and* $(stop == false)$ **do**
6: $\quad \Delta V_{ij} \leftarrow \Delta V_{ij} \cdot \frac{[\alpha(W_G \circ G)^T US + \gamma \Delta C_I]_{ij}}{\{\alpha[W_G \circ (US\Delta V^T)]^T US + \gamma \Delta V\}_{ij}}$
7: $\quad L \leftarrow \alpha \cdot \|W_G \circ (G - US\Delta V^T)\|_F^2 + \gamma \cdot \|\Delta V - \Delta C_I\|_F^2$;
8: $\quad$ **if** (L increases in this iteration) **then**
9: $\quad\quad stop = true$;
10: $\quad\quad$ Restore V' to its value in last iteration.
11: $\quad$ **end if**
12: **end while**
13: Append ΔC_I to $C_I \rightarrow C'_I$;
14: Append ΔV to $V \rightarrow V'$;
15: Calculate $US\Delta V^T \rightarrow G_r$;
16: Return $V', C'_I, Centroids'_V$, and G_r.

Data owner should hold the updated factor matrices (U', S, and V') and the cluster information (user/item cluster membership indicator matrices and centroids) for future update. Note that we leave the matrices S and V (S and U) unchanged in row update (column update), which does not indicate they will never change. We will show when Aux-NMF should be recomputed to ensure the data utility and privacy in the experimental study section.

V. Experimental Study

In this section, we discuss the test datasets, data preprocessing, evaluation strategy, and experimental results.

A. Data Description

In the experiments, we adopt MovieLens [24], Sushi [12] preference, and LibimSeTi [2] dating datasets as the test data. Table I collects the statistics of the datasets.

TABLE I: Statistics of the data

Dataset	#users	#items	#ratings	Sparsity
MovieLens	943	1,682	100,000	93.7%
Sushi	5,000	100	50,000	90%
LibimSeTi	2,000	5,625	129,281	98.85%

The public MovieLens dataset that we use has 943 users and 1,682 items. The 100,000 ratings, ranging from 1 to 5, were divided into two parts: the training set (80,000 ratings) and the test set (20,000 ratings). In addition to rating data, users' demographic information and items' genre information are also available.

The Sushi dataset describes users' preferences on different kinds of sushi. There are 5,000 users and 100 sushi items. Each user has rated 10 items, with a rating ranging from 1 to 5. That is to say, there are 50,000 ratings in this dataset. To build the test set and training set, for every user, we randomly select 2 out of 10 ratings and put them into the test set (10,000 ratings) while the rest of ratings are used as training set (40,000 ratings). Similar to MovieLens, the Sushi dataset comes with user's demographic information as well as item's group information and some attributes (e.g., the heaviness/oiliness in taste, how frequently the user eats the sushi etc.).

The LibimSeTi dating dataset is gathered by LibimSeTi.cz, an online dating website. It contains 17,359,346 anonymous ratings of 168,791 profiles made by 135,359 LibimSeTi users as dumped on April 4, 2006. However, only user's gender is provided with the data. We will show how to deal with this problem (lack of item information) in later section. Confined to the memory limitation of the test computer, we pick up 2,000 users and 5,625 items (profiles are considered as items for this dataset) with 108,281 ratings in training set and 21,000 ratings in test set. Ratings are on a $1 \sim 10$ scale where 10 is best.

B. Data Preprocessing

The proposed algorithms require user and item feature matrices as the input. To build such feature matrices, we pre-process the auxiliary information of users and items. In MovieLens dataset, user's demographic information includes user id, age, gender, occupation, and zip code. Amongst them, we utilize age, gender, and occupation as features. For age, the numbers are categorized into 7 groups: 1-17, 18-24, 25-34, 35-44, 45-49, 50-55, >=56. For gender, there are two possible values: male and female. As per statistics, there are 21 occupations: administrator, artist, doctor, and so on. Based on these possible values, we build a user feature matrix F_U with 30 features ($k_U = 30$), i.e., each user is represented as a row vector with 30 elements. An element will be set to 1 if the corresponding feature value is true for this user and 0 otherwise. An example is, for a 48 years old female user, who is an artist, the elements in the columns corresponding to female, 45-49, and artist should be set to 1. All other elements should be 0. Similar with user feature matrix, item feature matrix is built according to their genres. Movies in this dataset are attributed to 19 genres and hence the item feature matrix F_I has 19 features ($k_I = 19$) in it.

In Sushi dataset, we use some of the user's demographic information, i.e., gender and age. In this case, user's age has been divided into 6 groups by the data provider: 15-19, 20-29, 30-39, 40-49, 50-59, >=60. User gender consists of male and female, which is same as MovieLens data. Thus, the user feature matrix for this dataset has 5,000 rows and 8 columns. The item feature matrix, on the other hand, has 100 rows and 16 columns. The 16 features include 2 styles (maki and other),

2 major groups (seafood and other), and 12 minor groups (aomono (blue-skinned fish), akami (red meat fish), shiromi (white-meat fish), tare (something like baste; for eel or sea eel), clam or shell, squid or octopus, shrimp or crab , roe, other seafood, egg, meat other than fish, vegetables).

Different from MoiveLens and Sushi datasets, LibimSeTi dataset only provides user's gender as its auxiliary information so we directly use it as user's cluster indicator matrix C_U. It is worth noting that in this dataset, there are three possible gender values: male, female, and unknown. To be consistent, the number of user clusters is set to 3.

C. Evaluation Strategy

For comparison purposes, we run the proposed approach and the SVD-based data update approach [30] on the datasets to measure the error of unknown value imputation and the privacy level of the perturbed data, as well as their time cost. The SVD-based data update approach first uses the column mean to impute missing values in the new data and then performs the incremental SVD update on the imputed data. The machine we use is equipped with Intel® Core™ i5-2405S processor, 8GB RAM and is installed with UNIX operating system. The code was written and run in MATLAB.

We start with the partial training matrix R (also referred to as the original data, which is built by removing ratings left on some items or left by some users from the complete training matrix[3]), and then add the rest of data (also referred to as the new data) to R in several rounds.

When building R, we use the split ratio to decide how many ratings will be removed from the complete training data. For example, there are 1000 users and 500 items with their companion ratings in the training data. If the split ratio is 40% and we will do a row update, we use the first 400 rows as the original data, i.e., R ($\in \mathbb{R}^{400\times 500}$). The remaining 600 rows of the training matrix will be added to R in several rounds. Similarly, if we are going to perform a column update, we use the first 200 columns as the original data ($R \in \mathbb{R}^{1000\times 200}$) while the remaining 300 columns will be added to R in several rounds.

In each round, we add 100 rows/columns to the original data. If the number of the rows/columns of new data is not divisible by 100, the last round will update the rest. Therefore, in this example, the remaining 600 rows will be added to R in 6 rounds with 100 rows each. Note that Sushi data only has 100 items in total but we still want to test the column update on it so we add 10 items instead of 100 in each round.

The basic procedure of the experiments is as follows:

1) Perform Aux-NMF and SVD on R, producing the approximated matrix R_r (see Fig. 1);
2) Append the new data to R_r by iAux-NMF and SVD-based data update algorithm (SVDU for short) [30], yielding the updated rating matrix $\tilde{R}_r$;
3) Measure imputation error[4] and privacy of the updated rating matrix $\tilde{R}_r$;
4) Compare and study the results.

The imputation error is obtained by calculating the difference between the actual ratings in the test data and the imputed ratings in the released data. A common and popular criterion is the MAE (Mean Absolute Error), which can be calculated as follows:

$$MAE = \frac{1}{|TestSet|} \sum_{r_{ij}\in TestSet} |r_{ij} - p_{ij}| \qquad (38)$$

where r_{ij} is the actual value while p_{ij} is the predicted value.

When measuring the privacy, we define the privacy level in Definition 2

Definition 2: **Privacy level** $\Pi(Y|X)$ is a metric that indicates to what extent a random variable Y could be estimated if given random variable X.

$$\Pi(Y|X) = 2^{h(Y|X)} \qquad (39)$$

where $h(Y|X)$ is the differential entropy of Y given X.

This privacy measure was proposed by Agrawal et al.[1] and was applied to measure the privacy in collaborative filtering by Polat et al.[21], and Wang et al.[30]. In our experiment, we take $\Pi(Y|X)$ (the higher the better) as privacy measure to quantify the privacy, where random variable Y corresponds to the values in training set and X corresponds to the perturbed values (at same position as those in training set) in released data.

D. Results and Discussion

In this section, we present and discuss our experimental results in two stages. We first run Aux-NMF and SVD on the complete training data to evaluate the performance of the non-incremental algorithm. Then we follow the steps as specified in the previous section to evaluate the incremental algorithms.

1) Test on complete Training Data: Some parameters of the proposed algorithms need to be determined in advance. Table II gives the parameter setup in Aux-NMF (see Algorithm 1).

TABLE II: Parameter Setup in Aux-NMF

Dataset	α	β	γ	k	l	$MaxIter$
MovieLens	0.2	0	0.8	7	7	10
Sushi	0.4	0.6	0	7	5	10
LibimSeTi	1	0	0	3	10	10

For MovieLens dataset, we set $\alpha = 0.2$, $\beta = 0$, and $\gamma = 0.8$, which means that we rely mostly on the item cluster matrix, and then the rating matrix, whereas eliminate the user cluster matrix. This combination was selected after probing many possible cases. We will discuss how we choose the parameters in Section V-D3. We believe there still exist better combinations. Both k and l are set to 7. We set these values because K-Means was prone to generate empty clusters with

[3]Here, "complete" means all the ratings from the dataset are in the matrix. It is still a sparse matrix.

[4]We use the term "imputation error" because all missing values are imputed and will be compared with the real values, though no specific imputation technique is used in Aux-NMF and iAux-NMF.

greater k and l, especially on the data with very few users or items. Note that if β or γ is a non-zero value, the user or item cluster matrix will be used and k or l is equal to the number of user clusters or item clusters. As long as β or γ is zero, the algorithm will eliminate the corresponding cluster matrix and k or l will have nothing to do with the number of user clusters or item clusters.

For Sushi dataset, we set $\alpha = 0.4$, $\beta = 0.6$, and $\gamma = 0$. The parameters indicate that the user cluster matrix plays the most critical role during the update process. In contrast, rating matrix is the second important factor as it indicates the user preference on items. The item cluster matrix seems trivial so it does not participate the computation. We set k to 7 and l to 5 based on the same reason as mentioned in previous paragraph.

For LibimSeTi dataset, we give the full weight to the rating matrix. Zero weight is received for user and item cluster matrices since they do not contribute anything to the good results. As mentioned in data description, user's auxiliary information only includes the gender with three possible values. So we set k to 3. In this case, l only denotes the column rank of V and is set to 10.

In SVD, since it cannot run on an incomplete matrix, we use item mean to impute the missing values (see [30]). The rank is set to 13 for MovieLens, 7 for Sushi, and 10 for LibimSeTi. Table III presents the results on three datasets.

TABLE III: Results on MovieLens dataset

Dataset	Method	MAE	$\Pi(Y\|X)$	Time Cost
MovieLens	Aux-NMF	0.7481	1.2948	0.9902s
	SVD	0.7769	1.2899	34.1341s
Sushi	Aux-NMF	0.9016	1.4588	0.5350s
	SVD	0.9492	1.4420	5.4175s
LibimSeTi	Aux-NMF	1.2311	1.0715	5.7962s
	SVD	1.2154	1.0537	390.2246s

In this table, the time cost of SVD includes the imputation time while the time cost of Aux-NMF includes the clustering time. For instance, on MovieLens dataset, the imputation took 32.2918 seconds and SVD itself took 1.8423 seconds, as 34.1341 seconds in total; the clustering time took 0.0212 seconds and Aux-NMF itself took 0.9690 seconds, as 0.9902 seconds in total. As can be seen, Aux-NMF outperformed SVD in all aspects on all three datasets. We notice that the former ran much faster than the latter (saves 97% time on MovieLens, 90% time on Sushi, and 98% time on LibimSeTi). This is mainly because SVD-based algorithm needs imputation, which is time consuming, but for Aux-NMF, it can directly work on sparse matrix though it needs to cluster beforehand (it is very fast in general).

It is interesting to take a look at the results of running K-Means on the final matrix generated by Aux-NMF and the matrix generated by SVD. As shown in Fig. 2(a), MovieLens users with ratings produced by Aux-NMF were clustered into 7 groups with clear boundaries. However, the result is different for SVD - most users were grouped together and thus the clusters cannot be distinguished from each others. Note that the axes in both figures are ratings left by users on items. The results indicate the more normally distributed ratings in the matrix generated by Aux-NMF than SVD. Remember that our goal is to provide good imputation accuracy as well as high privacy level. In addition, the data should look as if it is the real data. To this end, we should make the ratings distribute normally, i.e., people may leave more 3 stars on a $1 \sim 5$ scale than 1 star and 5 stars. In this regard, Aux-NMF generated more reasonable data than SVD did.

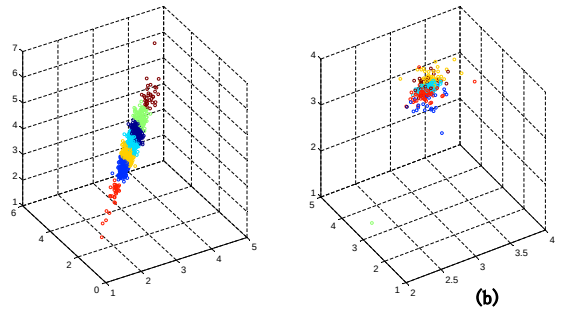


Fig. 2: Clustering results on ratings predicted by Aux-NMF (a) and SVD (b) on MovieLens dataset

2) The Incremental Case: In previous section, we examined the Aux-NMF on three datasets in terms of MAE, privacy level, as well as time cost. Now we measure the same metrics on iAux-NMF (incremental Aux-NMF).

Fig. 3 shows the time cost for updating new rows and columns by iAux-NMF and SVDU (SVD-based data update algorithm). We use "RowN" and "ColumnN" to represent row and column updates in iAux-NMF. Similarly, "RowS" and "ColumnS" are for row and column updates in SVDU. We use the same parameter setup in Table II.

It can be seen that iAux-NMF outperformed SVDU in both row and column updates. As pointed out in Section IV-B, the time complexity of row update in iAux-NMF is $O(pn(l+k)$ and column update has a time complexity of $O(qm(l+k)$. As a reference, the time complexities of row and column updates in SVDU are $O(k^3 + (m+n)k^2 + (m+n)kp + p^3)$ and $O(k^3 + (m+n)k^2 + (m+n)kq + q^3)$, respectively. When the rating matrix has high dimensions, the time cost difference can be huge. For example, the LibimSeTi dataset has both more users and more items than MovieLens so the improvement of iAux-NMF over SVDU plotted in Fig. 3(c) was greater than Fig. 3(a). However, the Sushi data is a bit special as the time difference between two methods in row update was very small, though iAux-NMF still ran faster. In Section V-D1, we broke the time cost of both methods into two pieces: for SVDU, the time consists of imputation time and SVD computation time; for Aux-NMF, the time consists of clustering time and Aux-NMF computation time (Before running the algorithms, the parameters need to be determined. We will discuss the time cost for this part in Section V-D3.). By tracking the time cost of each stage, we found that the imputation in SVDU took considerably shorter time in row update than column update on this dataset but the time cost of Aux-NMF in row update and column update did not differ a lot. Essentially, the faster imputation in row update can be attributed to the small number of items. Since SVDU uses the column mean to impute the

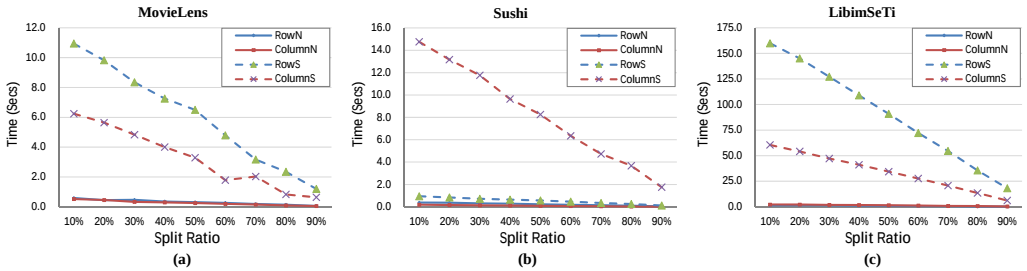


Fig. 3: Time cost variation with split ratio

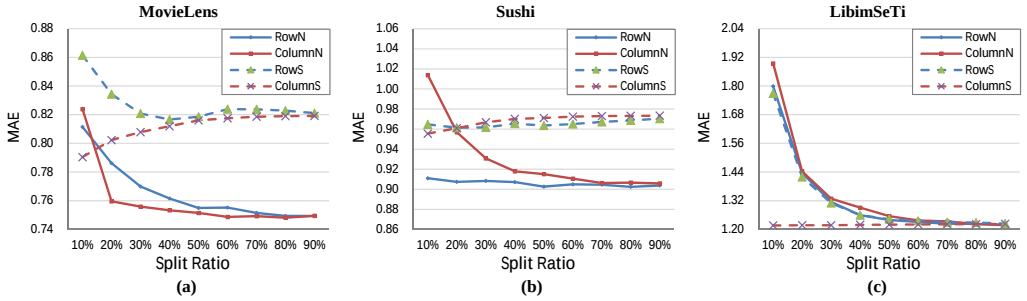


Fig. 4: MAE variation with split ratio

missing values, if there are only a few items, the mean value calculation can be fast.

However, with the substantial improvement in time cost, iAux-NMF should not produce a significantly higher imputation error than SVDU.

Fig. 4 shows the mean absolute errors of the prediction. When the split ratio was greater than 20%, iAux-NMF achieved lower errors than SVUD on MovieLens and Sushi datasets. The average improvement on MovieLens was 9.79% for row update and 9.76% for column update. The Sushi dataset had a little less average improvement than MoiveLens but it was still noticeable. Nevertheless, both of them had large errors by iAux-NMF than by SVD when the split ratio was less than 20%. This is because the centroids picked up by K-Means algorithm did not distribute over the data that was not large enough to reflect the global picture. With badly selected centroids, K-Means cannot produce a good clustering result which further affects the Aux-NMF and iAux-NMF so the errors would be large. Unlike MovieLens and Sushi, the LibimSeTi dataset got different results. In this case, iAux-NMF still performed better than SVDU but the gap tended to be smaller as the split ratio increased. The results imply that auxiliary information is important to iAux-NMF as it is used as constraint in the update process. On the contrary, SVDU does not need it. This can explain why SVDU performed better than iAux-NMF on LibimSeTi (no auxiliary information is used).

In Section IV-B2, we mentioned the issue of Aux-NMF re-computation. As presented in Fig. 4, the MAE's of both row and column updates on MovieLens dataset dropped more slowly at 70% and nearly kept the same after this point. Similarly but more interestingly, the MAE of row update on Sushi dataset began to increase at 70%. Therefore, a re-computation can be performed at 70% for these two datasets. For LibimSeTi dataset, the MAE's did not seem to stop decreasing so the re-computation is not immediately necessary.

In addition to MAE, we want to investigate the privacy metrics presented in Section V-C. The privacy level with varying split ratio is plotted in Fig. 5. The curve shows that the privacy level of the data produced by iAux-NMF were higher and more stable than SVDU while the latter had decreasing trend with greater split ratios. The results are encouraging.

As a summary, the iAux-NMF data update algorithm ran much faster than SVDU while maintaining nearly the same data utility and privacy as SVDU, if not better.

3) Parameter Study: In iAux-NMF, three parameters, i.e., α, β, and γ need to be set. In this section, we do some comparisons over several parameter combinations and discuss the results. Note that we keep the split ratio at 40% and pre-generate the initial random matrices in Algorithms 2 and 3 to eliminate the effect of randomness in the experiments. We adopt the parameter setup in Table II because it is the best combination obtained by probing many possible cases. The pseudocode in Algorithm 4 shows the procedure to find out the parameters that produce the lowest MAE's. The step is set to 0.1 when we increment the parameters. Since there is a constraint $\alpha + \beta + \gamma = 1$, the total number of parameter combinations is 66. It took 806.28 seconds to run a full test on MovieLens dataset, 1116.9 seconds on Sushi, and 11517.87 seconds on LibimSeTi. The times are relatively long when compared with the times of running the incremental algorithms. However, the parameters only need to be determined offline once so it will not affect the online performance.

Table IV lists some representative combinations with their

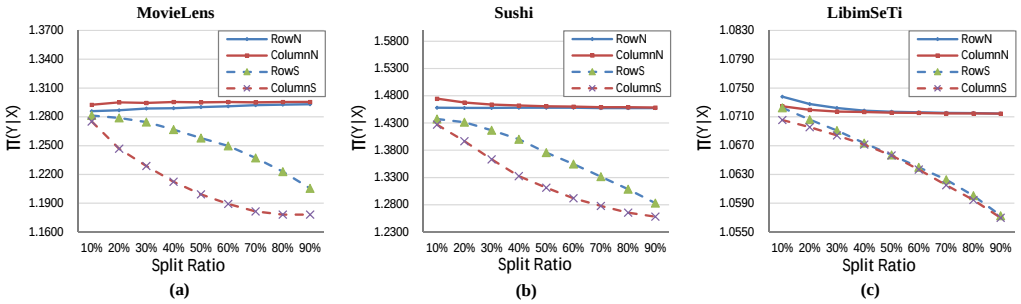


Fig. 5: Privacy level variation with split ratio

Algorithm 4 Pseudocode for Parameter Probing

1: **for** $\alpha = 0 : 0.1 : 1$ **do**
2: **for** $\beta = 0 : 0.1 : 1 - \alpha$ **do**
3: $\gamma = 1 - \alpha - \beta$.
4: Run Aux-NMF and iAux-NMF on a dataset with parameter α, β, and γ, saving the MAE's as well as α, β, and γ to the corresponding variables.
5: **end for**
6: **end for**
7: Find out the lowest MAE and obtain the associated parameters.

results on MovieLens dataset. The best combinations are in bold font. We notice that if the updates simply relied on the rating matrix, the results were only a little worse than taking into account the auxiliary information. In contrast, if only the auxiliary information was utilized, the MAE was unacceptable, though the privacy level was the highest. It is clear that between user features and item features, the latter made good contribution to the results while the former seems trivial. Nevertheless, the weight of rating matrix can be lowered but should not be removed. The Sushi dataset (Table V) had a similar conclusion but it is the user features that played a more dominant role.

TABLE IV: Parameter Probe on MovieLens dataset

Parameters	Update	MAE	$\Pi(Y\|X)$
$\alpha = 1, \beta = 0, \gamma = 0$	Row	0.7643	1.2913
	Column	0.7538	1.2964
$\alpha = 0.5, \beta = 0.5, \gamma = 0$	Row	0.7643	1.2913
	Column	0.7539	1.2963
$\alpha = 0.5, \beta = 0, \gamma = 0.5$	Row	0.7624	1.2909
	Column	0.7534	1.2958
$\alpha = 0, \beta = 0.5, \gamma = 0.5$	Row	0.9235	1.3149
	Column	0.9164	1.3150
$\boldsymbol{\alpha = 0.2, \beta = 0, \gamma = 0.8}$	Row	0.7616	1.2890
$\boldsymbol{\alpha = 0.4, \beta = 0, \gamma = 0.6}$	Column	0.7533	1.2955

As shown in Table VI, the rating matrix of LibimSeTi dataset was the only information used in the computation. This indicates that even the dataset comes with users' genders, they did not help in our model. This is reasonable as the gender is not a necessary factor for people to determine their ratings (A female can rate another female with a fairly high rating.). Note that since there is no item features coming with this dataset, γ was always set to zero.

TABLE V: Parameter Probe on Sushi dataset

Parameters	Update	MAE	$\Pi(Y\|X)$
$\alpha = 1, \beta = 0, \gamma = 0$	Row	0.9083	1.4578
	Column	0.9221	1.4613
$\alpha = 0.5, \beta = 0.5, \gamma = 0$	Row	0.9073	1.4580
	Column	0.9201	1.4614
$\alpha = 0.5, \beta = 0, \gamma = 0.5$	Row	0.9085	1.4580
	Column	0.9221	1.4614
$\alpha = 0, \beta = 0.5, \gamma = 0.5$	Row	1.0468	1.4851
	Column	1.0371	1.4849
$\boldsymbol{\alpha = 0.4, \beta = 0.6, \gamma = 0}$	Row	0.9071	1.4580
$\boldsymbol{\alpha = 0.2, \beta = 0.8, \gamma = 0}$	Column	0.9180	1.4620

Therefore, we can conclude that, the rating matrix should always be utilized while the auxiliary information makes contributions to the improved results as well.

TABLE VI: Parameter Probe on LibimSeTi dataset

Parameters	Update	MAE	$\Pi(Y\|X)$
$\boldsymbol{\alpha = 1, \beta = 0, \gamma = 0}$	Row	1.2589	1.0719
	Column	1.2911	1.0717
$\alpha = 0.5, \beta = 0.5, \gamma = 0$	Row	1.3378	1.0713
	Column	1.3926	1.0709
$\alpha = 0, \beta = 1, \gamma = 0$	Row	5.4017	1.0782
	Column	5.4017	1.0782

VI. Conclusion and Future Work

In this paper, we proposed a NMF-based privacy preserving data update approach for collaborative filtering purpose. This approach utilizes the auxiliary information to build the cluster membership indicator matrices for users and items. These matrices are regarded as additional constraints in updating the weighted nonnegative matrix tri-factorization. The proposed approach, named iAux-NMF, can incorporate the incremental data into existing data quite efficiently while maintaining the high data utility and privacy. Furthermore, the inevitable missing value imputation issues in collaborative filtering is solved in a subtle manner by this approach without using any particular imputation methods. Experiments conducted on three different datasets demonstrate the superiority of iAux-NMF over the existing privacy-preserving SVD-based data update method in the situation of incremental data update.

In future work, we will consider the automated clustering update when new data comes in. This new feature will decide the number of clusters by itself and recompute the NMF when needed. We believe it can provide better data utility and privacy. We will also investigate the distributed data update in collaborative filtering and attempt to propose the corresponding distributed algorithms.

VII. Acknowledgments

We thank Oldrich Neuberger for providing the LibimSeTi dating data and Lukas Brozovsky for cleaning up and generating the dataset.

References

[1] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '01, pages 247–255. ACM, 2001.

[2] L. Brozovsky and V. Petricek. Recommender system for online dating service. In *Proceedings of Znalosti 2007 Conference*. VSB, 2007.

[3] J. Canny. Collaborative filtering with privacy. In *Proceedings of the 2002 IEEE Symposium on Security and Privacy*, pages 45–57. IEEE Computer Society, 2002.

[4] G. Chen, F. Wang, and C. Zhang. Collaborative filtering using orthogonal nonnegative matrix tri-factorization. In *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops*, ICDMW '07, pages 303–308. IEEE, 2007.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[6] C. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, SDM '05, pages 606–610. SIAM, 2005.

[7] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, SIGKDD '06, pages 126–135. ACM, 2006.

[8] Eurostat. *Manual on Disclosure Control Methods*. Office for Official Publications of the European Communities, 1996.

[9] S. Ferdowsi, V. Abolghasemi, and S. Sanei. A constrained nmf algorithm for bold detection in fmri. In *Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 77–82, 2010.

[10] P. O. Hoyer and P. Dayan. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:14571469, 2004.

[11] S. M. A. Kabir, A. M. Youssef, and A. K. Elhakeem. On data distortion for privacy preserving data mining. In *Proceedings of Canadian Conference on Electrical and Computer Engineering*, CCECE 2007, pages 308 – 311. IEEE, 2007.

[12] T. Kamishima and S. Akaho. Efficient clustering for orders. In *Proceedings of the 2nd International Workshop on Mining Complex Data*, pages 274–278, 2006.

[13] J. Kim and H. Park. Sparse nonnegative matrix factorization for clustering. Technical report, , Georgia Institute of Technology, 2008.

[14] H. Kuhn and A. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, 1951.

[15] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2001.

[16] H. Li, T. Adali, W. Wang, D. Emge, and A. Cichocki. Non-negative matrix factorization with orthogonality constraints and its application to raman spectroscopy. *Journal of VLSI Signal Processing Systems*, 48(1-2):83–97, 2007.

[17] H. Liu and Z. Wu. Non-negative matrix factorization with constraints. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI '10, pages 506–511. AAAI, 2010.

[18] Y. Mao and L. K. Saul. Modeling distances in large-scale networks by matrix factorization. In *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, IMC '04, pages 278–287. ACM, 2004.

[19] F. McSherry and I. Mironov. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 627–636. ACM, 2009.

[20] V. P. Pauca, F. Shahnaz, M. W. Berry, and R. J. Plemmons. Text mining using nonnegative matrix factorizations. In *Proceedings of 2004 SIAM Interational Conference on Data Mining*, volume 54 of *SDM '09*, pages 452–456. SIAM, 2004.

[21] H. Polat and W. Du. Privacy-preserving collaborative filtering. *International Journal of Electronic Commerce*, 9(4):9–35, 2005.

[22] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, pages 175–186. ACM, 1994.

[23] R. Sandler and M. Lindenbaum. Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1590–1602, 2011.

[24] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. T. Riedl. Application of dimensionality reduction in recommender systems – a case study. In *Proceedings of ACM WebKDD Workshop*. ACM, 2000.

[25] N. Thapa, L. Liu, P. Lin, J. Wang, and J. Zhang. Constrained nonnegative matrix factorization for data privacy. In *Proceedings of the 7th International Conference on Data Mining*, DMIN '11, pages 88–93, 2011.

[26] J. Tougas and R. J. Spiteri. Updating the partial singular value decomposition in latent semantic indexing. *Computational Statistics & Data Analysis*, 52:174–183, 2007.

[27] S. Vucetic and Z. Obradovic. Collaborative filtering using a regression-based approach. *Knowledge and Information Systems*, 7:1–22, 2005.

[28] J. Wang, J. Zhan, and J. Zhang. Towards real-time performance of data value hiding for frequent data updates. In *Proceedings of the IEEE International Conference on Granular Computing*, pages 606–611. IEEE Computer Society, 2008.

[29] J. Wang, W. Zhong, and C. Zhang. Nnmf-based factorization techniques for high-accuracy privacy protection on non-negative-valued datasets. In *Proceedings of the sixth IEEE International Conference on Data Mining Workshops*, ICDM Workshops 2006, pages 513–517. IEEE, 2006.

[30] X. Wang and J. Zhang. SVD-based privacy preserving data updating in collaborative filtering. In *Proceedings of the World Congress on Engineering 2012*, WCE 2012, pages 377–284. IAENG, 2012.

[31] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 267–273. ACM, 2003.

[32] K. Yu, S. Zhu, J. Lafferty, and Y. Gong. Fast nonparametric matrix factorization for large-scale collaborative filtering. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 211–218. ACM, 2009.

[33] J. Zhang. Image fusion based on nonnegative matrix factorization. In *Proceedings of 2004 International Conference on Image Processing*, volume 2 of *ICIP '04*, pages 973–976. IEEE, 2004.

[34] S. Zhang, W. Wang, J. Ford, and F. Makedon. Learning from incomplete ratings using non-negative matrix factorization. In *Proceedings of the Sixth SIAM International Conference on Data Mining*, pages 548–552. SIAM, 2006.

Surface Texture Synthesis and Mixing Using Differential Colors

Qing Wu Lin Shi Stephen Bond Yizhou Yu

University of Illinois at Urbana-Champaign
Email: {linshi,qingwu1,sdbond,yyz}@uiuc.edu

***Abstract*—In neighborhood-based texture synthesis, adjacent local regions need to satisfy color continuity constraints in order to avoid visible seams. Such continuity constraints seriously restrict the variability of synthesized textures, making it impossible to generate new textures by mixing multiple input textures with very different base colors. In this paper, we propose to relax such restrictions and decompose synthesis into two relatively disjoint stages. In the first stage, an intermediate synthesized texture is generated by only considering the high frequency details during region search and matching. Such a scheme broadens the search space during texture synthesis, but may produce obvious seams due to large discontinuities in low frequency components. In the second stage, instead of performing local feathering along these discontinuities, we perform Laplacian texture reconstruction, which retains the high frequency details but computes new consistent low frequency components to eliminate the seams. It does not only affect texels close to the discontinuities, but also modifies the rest of the texels. Therefore, it can be viewed as a global feature-preserving smoothing step, and is more effective than local feathering. Experiments indicate that our two-stage synthesis can produce desirable results for regular texture synthesis as well as texture mixing from multiple sources.**

I. Introduction

Texture synthesis has been widely recognized as an important research topic in computer graphics. Early texture synthesis algorithms were based on global statistical models. Instead of enforcing global statistics, preserving the local arrangement of pixels has proven to be more effective in terms of visual quality. This intuition led to the neighborhood-based search-and-copy algorithms. Given a small texture example, these algorithms can produce larger textures that have similar texture elements and structures as the given example. Every output texture from such neighborhood-based texture synthesis is essentially a spatial rearrangement of the original local regions in the given example. When there is only a single small texture example, the number of possible rearrangements is actually limited because adjacent local regions need to satisfy continuity restrictions.

We propose to relax such neighborhood-based texture synthesis along two directions to improve the variability of the synthesized results without compromising their visual quality. First, relax the continuity restrictions. Previously, adjacent local regions in the output texture are typically required to have pixelwise color similarity in their overlapping portion. When a new local region needs to be chosen from the texture example, such a stringent condition results in a very small number of candidates and quite often zero candidates. We suggest to relax such region matching by focusing on the high frequency components only and overlooking the average color and intensity. Indeed, it is the high frequency components that play the most important role in characterizing a texture.

Second, allow multiple input texture examples. Sampling local regions from a single small texture example can only produce very limited variability. Ideally, example-based synthesis should generate results by sampling a large database. However, different textures may be acquired by different imaging devices and/or settings, under different illumination conditions, etc. Even the same real-world texture, such as grass, can appear very different in different texture images. Effectively sampling, matching and mixing local regions from multiple texture examples simultaneously is nontrivial.

In this paper, we focus on surface texture synthesis and propose a novel two-stage synthesis approach to accommodate these two relaxations. In the first stage, an intermediate synthesized texture is generated by only considering the high frequency details during region search and matching. It is achieved by using both features and rectified versions of the input textures. Each pixel value in the rectified textures is defined by its original value normalized by the accumulated intensity within a neighborhood. Such a scheme broadens the search space during texture synthesis. It facilitates sampling local regions from different texture examples as well as placing regions with very different average colors and intensities next to each other in the output texture. However, this intermediate texture has obvious seams due to large discontinuities in low frequency components.

In the second stage, to seamlessly mix local regions together and create smooth transitions among them, we perform texture reconstruction using differential colors, which is in the same spirit of Poisson image editing [1] and related surface editing [2]. It retains the high frequency details, but computes new consistent low frequency components for all local regions so that the seams among them disappear. Therefore, it can be viewed as a global feature-preserving smoothing step, and is more effective than local feathering. The Laplacian operator extracts local differential quantities which represent high frequency texture details. Given the Laplacian of the intermediate texture, we set up a sparse linear system with the new texture colors as the unknowns. The solution of this system not only retains the original texture details, but also provides a consistent coloring to all the texels in the synthesized texture without discontinuities along the boundaries of adjacent regions.

A. Related Work

This paper is partially inspired by the recent success of texture synthesis [3], [4], [5], [6], [7], [8], [9], [10], [11], [12]. 2D textures are frequently modeled as Markov Random Fields, which give rise to the neighborhood based texture synthesis algorithms. Some algorithms model textures as a set of features, and generate new images by matching feature statistics, such as histograms and co-occurrences, potentially across multiple resolutions [3]. Patch-based texture synthesis [4], [5], [7], [8] achieved better results than earlier methods in terms of both quality and efficiency. Feature continuity on the boundary of two adjacent patches is an important issue and [5], [7], [8] have attempted to alleviate this problem using dynamic programming, graph cuts, and feature maps, respectively. Further improvements include real-time parallel synthesis on GPUs [9] and multiscale synthesis [10]. Recently, techniques have been developed to support the synthesis of textures with multiple layers using either signed distance functions or levelsets [11], [12].

There has also been much work [13], [14], [15], [16], [17], [18] on generalizing 2D texture synthesis onto meshes with arbitrary topology. Neighborhood based 2D synthesis was generalized to meshes in [13], [14], which perform hierarchical vertex-based synthesis. A binary texton mask was introduced in [16] as guidance data to improve synthesis results and reduce the number of broken features. Patch-based synthesis has also been generalized to meshes [15], [17]. A hierarchical patch-based approach was proposed in [15], and an efficient synthesis method on triangle meshes was introduced in [17] . A very fast synthesis technique accelerated by precomputed Jump Maps was introduced in [18]. Neighborhood or patch matching in these methods is directly based on color differences instead of differences in high frequency components.

Meanwhile, researchers have synthesized textures from multiple input examples by mixing together different elements from them [3], [16]. In [3], statistical learning trees are used to mix textures. On the other hand, the technique in [16] focuses on creating a progressive transition between different texture elements. However, these texture mixing techniques cannot globally adjust the colors of the mixed textures to make them more consistent with each other.

II. Overview

The input to our algorithm is a triangle mesh as well as one or more texture examples. The output is the same mesh covered with a texture synthesized from the given examples. In the case of multiple input textures, the synthesized texture is a spatial mixture of the texture elements from the texture examples.

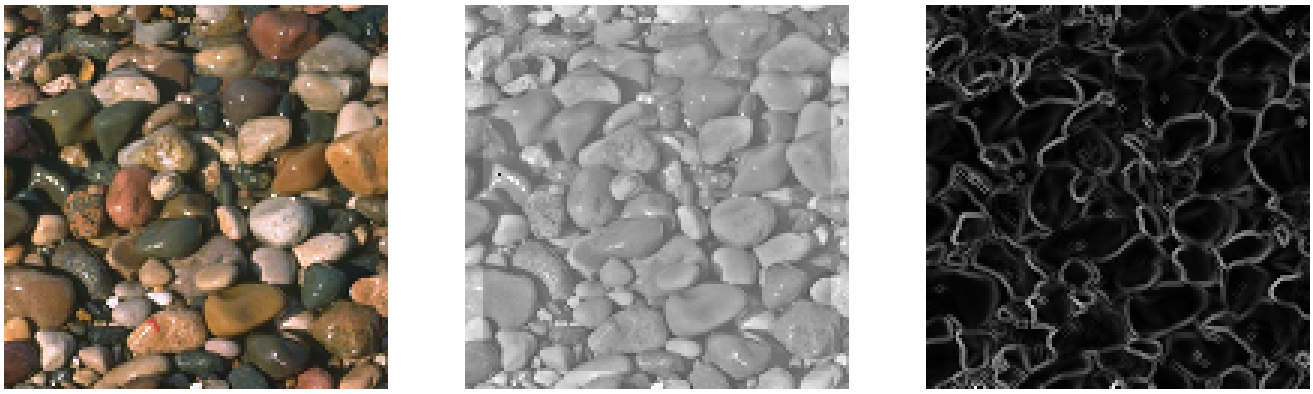
Our Laplacian texture synthesis algorithm has three basic steps.

1) Apply a revised version of the method in [17] to generate an initial texture patch assignment on the mesh. The method in [17] does not directly synthesize textures on a mesh. Instead, it assigns a triangular texture patch in the input textures to each triangle in the mesh. The assigned texture patches of two adjacent triangles have a certain degree of continuity along their shared edge. Our revised version tries to emphasize high frequency details, but overlook the differences in the average colors of local regions. At the end of this step, each triangle in the mesh is associated with three pairs of texture coordinates which record the locations of the corners of its corresponding triangular texture patch in the input texture examples.
2) Each triangle in the mesh is tesselated with a high-resolution grid and the assigned texture patch of the triangle is resampled onto this grid. A graphcut algorithm is further executed to refine the boundary between two adjacent texture patches so that such a boundary is not necessarily coincidental with the shared edge between two adjacent triangles any more. As a result, the transitions of details among texture patches are improved though their average colors may still be quite different.
3) Laplacian texture reconstruction is performed simultaneously on all the resampled texture patches from the previous step to eliminate the color discontinuities between adjacent patches. The Laplacian at each grid point is obtained from the original colors in the input texture examples.

III. Initial Texture Assignment

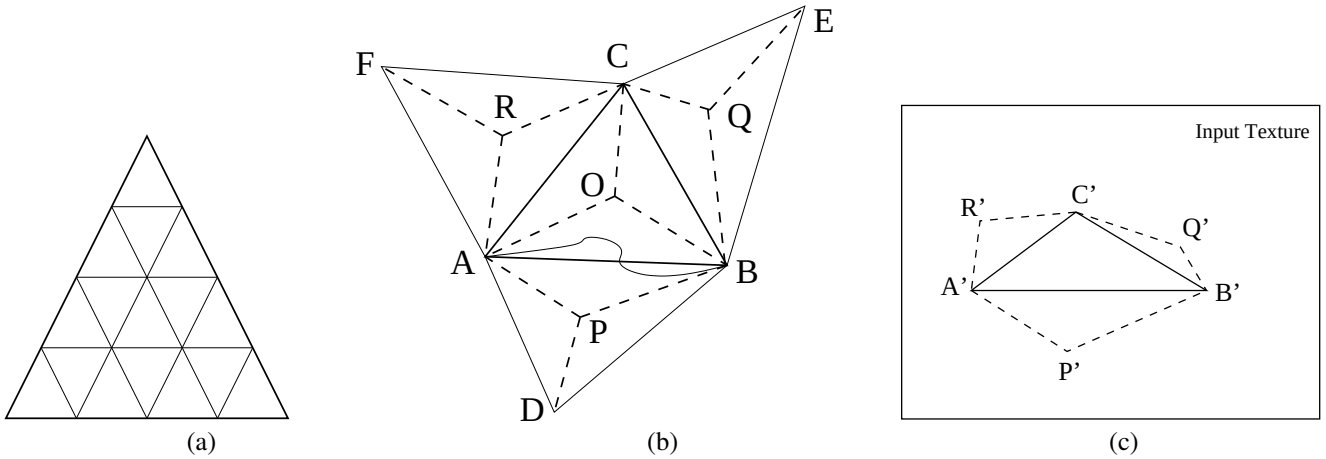
Our initial texture assignment is based on the method in [17], where a texton is defined to be a distinct local texture neighborhood. By clustering all neighborhoods with a fixed size from the given texture example, a small collection of textons can be extracted. They are the representatives of the clusters. During synthesis, triangular texture patches are grown on the mesh one by one to cover all the triangles. Note that each triangle in the mesh shares an edge with at most three adjacent triangles. During each step of synthesis, this method focuses on one triangle and counts the number of its adjacent triangles that have been covered with texture. If none of them has been covered, the current triangle is a seed and should be covered with a random patch. If one of them has been covered, we need to search for a texture patch in the given texture example that agrees well with the texture on this adjacent patch, which means that the two texton sequences on the shared edge should be similar. This strategy can be easily generalized to cases where two or three adjacent triangles are already covered with textures.

To emphasize high frequency details but overlook differences in average intensity and color, in our revised version of this method, we utilize a rectified version of each input texture. The rectified version of a texture is a greyscale image with a normalized intensity value at each pixel. We set the initial greyscale value at a pixel to be its luminance value which is a weighted average of the original three color channels. Suppose we define an $N \times N$ neighborhood for each pixel and the luminance at a neighbor (i, j) is I_{ij}. The luminance value at each pixel is further normalized by the accumulated luminance in its neighborhood, which is $(\sum_{ij} I_{ij}^2)^{1/2}$. In this rectified texture, we essentially have removed the low frequency components, but retained the important high frequency details. We only use greyscale values because they are weighted averages of three color channels and contain high frequency details from all of them. In practice, using greyscale values indeed can



(a) (b) (c)

Fig. 1. (a) The original PEBBLES texture. (b) The rectified greyscale image of (a). (c) A feature image of (a) obtained by filtering.



(a) (b) (c)

Fig. 2. (a) A fine grid defined within a triangle. (b) A graphcut is performed inside the quadrilateral region $OAPB$ to obtain a refined boundary between the two adjacent texture patches. (c) An extended hexagonal texture patch corresponding to ΔABC in (b).

produce better matching results than using three independent color channels. An example of a rectified texture is shown in Fig. 1(b). In practice, we set the size of the neighborhoods to be 11×11.

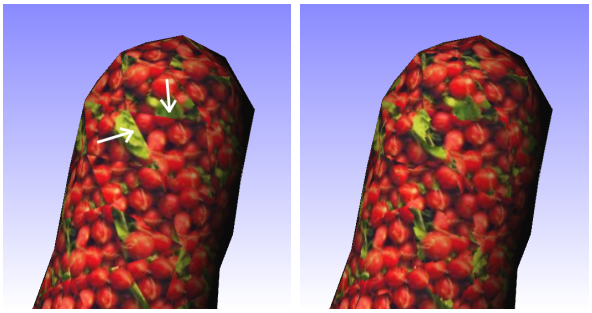
In addition to the rectified textures, we also obtain a feature image for each of the input texture examples. We first apply bilateral filtering [19] to remove noise while preserving features. In the bilateral filter, the scale of the closeness function σ_d is set to 2.0, and the scale of the similarity function σ_r is set to 10 out of 256 greyscale levels. We then use finite differences along the two image axes as a simple gradient estimator to obtain an edge response at every pixel. The pixelwise gradient estimation is used to form the feature images. An example of a feature image is shown in Fig. 1(c).

In our revised version of the method in [17], we use these rectified textures along with the feature images as the input to texton clustering. Thus, the neighborhood corresponding to each texton has a normalized greyscale pattern and a feature pattern. Both patterns emphasize high frequency details. In practice, weighted versions of these patterns are used for texton clustering. The weight for the greyscale pattern is set to 1.0, and the weight for the feature pattern is set to 0.3. These weighted patterns are treated as different channels of the same texture neighborhood during texton clustering. Once we have the collection of textons, the rest of the synthesis steps follow [17].

When there are multiple input textures, every time we need to search for a texture patch for a triangle, we find the best candidate from each input texture and then choose among them the one with the highest matching score. Usually, we would like to set up for each input texture a target percentage in the output texture. To approximately control the synthesis process using these target percentages, we define a Gaussian function for each input. The standard deviation of the Gaussian is set to be the target percentage of the input texture. The function value of the Gaussian is used to modulate the matching score. When the actual percentage is lower than the target percentage, the Gaussian returns a large value which does not obviously affect the matching score; when the actual percentage is higher than the target percentage, the Gaussian returns a small value which significantly decreases the matching score. Thus, these Gaussian functions implicitly control the likelihood of sampling a specific input texture.

IV. Texture Resampling and Boundary Refinement

Since we would like to perform boundary refinement on the triangular texture patches and Laplacian reconstruction over



(a) (b)

Fig. 3. (a) Initial texture patch assignment result with two of the seams indicated by arrows. (b) Result obtained from boundary refinement with Graphcut.

the entire synthesized texture, indexing the texture patch for each triangle as three pairs of texture coordinates in the input texture space becomes insufficient. To facilitate these later steps, we resample the texture patches onto a high-resolution grid over the original mesh surface. The high-resolution grid within each triangle is set up using barycentric coordinates as shown in Fig. 2(a). That is, every edge of the triangle has the same number of sample points. We actually further enforce that all edges in the triangle mesh have the same number of sample points. Thus, the subgrids within two adjacent triangles coincide on their shared edge to avoid T-junctions. If the original triangle mesh has some overly large or elongated triangles, we split those triangles in a preprocessing step while avoiding T-junctions.

We resample the texture patches previously assigned to the triangles onto this high-resolution grid. As shown in Fig. 2(b)-(c), suppose a triangle ΔABC in the mesh has a corresponding triangular texture patch $\Delta A'B'C'$ in one of the input texture examples. To facilitate boundary refinement at a later step, we actually resample an area larger than $\Delta A'B'C'$. Suppose ΔABD, ΔBCE and ΔCAF are the three triangles adjacent to ΔABC. Their centers are P, Q and R, respectively. We first flatten these three triangles onto the same plane where ΔABC resides and obtain the new locations of their centers. From these new locations, we can further obtain their corresponding locations P', Q' and R' in the 2D texture space. We resample the entire hexagonal area $A'P'B'Q'C'R'$ in the texture space onto the corresponding region, $APBQCR$, of the high-resolution grid. Thus, each resampled texture patch of a triangle extends into its three adjacent triangles. Suppose the center of ΔABC is O. During this resampling, ΔOAB, ΔOBC and ΔOCA not only obtain color values from the texture patch originally assigned to ΔABC, but also obtain a second color value from the extended hexagonal patches corresponding to the three adjacent triangles of ΔABC.

During initial texture patch assignment discussed in the previous section, each triangle is assigned a triangular texture patch. The boundary between two adjacent texture patches coincide with the shared edge of their corresponding triangles. In a subsequent boundary refinement procedure, we apply the graphcut algorithm in [7] to refine the boundaries between resampled texture patches on the high-resolution grid. We need to take into account the extended hexagonal texture patches to perform this procedure. Consider triangles ΔABC and ΔABD in Fig. 2(b). Suppose their hexagonal texture patches are HTP_O and HTP_P, respectively. These two texture patches have an overlapping quadrilateral region, $OAPB$. We set up a minimum graph cut problem as follows. The grid points closest to OA and OB are constrained to have colors from the texture patch HTP_O while the grid points closest to PA and PB are constrained to have colors from the patch HTP_P. The vertices A and B also have fixed colors. We then seek a minimum graph cut between A and B and within the region $OAPB$. The algorithm in [7] is applied to find this cut which can provide a better transition between the high frequency details of the two texture patches than the original boundary. The grid points falling on the same side of the cut as O obtain colors from patch HTP_O while the grid points on the other side of the cut obtain colors from patch HTP_P. Note that we still use the rectified textures during boundary refinement because the original texture colors may have large discontinuities along the triangle edges, which prevent the graphcut algorithm to find a different cut that provides better transition for high frequency details. Fig. 3(a)-(b) demonstrate the effectiveness of this boundary refinement procedure.

Since the colors of the mesh vertices are not refined at all during the aforementioned boundary refinement, we also designed another graphcut procedure specifically tailored for them. We first define an umbrella region centered at each vertex, and then flatten that region onto a parameterization plane. A subsequent graph cut is performed in this flattened region to refine the boundaries of the texture patches there. However, in our experiments, we have not observed any obvious improvements in visual quality due to this vertex-centric refinement. Therefore, we leave it as an optional step.

V. Laplacian Texture Reconstruction

The synthesis process in the previous sections focuses on high frequency details. We call the surface texture synthesized by the previous steps the *intermediate* texture. There are obvious seams inbetween adjacent texture patches in the intermediate texture because of discontinuities in the low frequency components. To remove these large discontinuities while still preserving high frequency texture details, we present a texture reconstruction technique based on the Laplacian operator which encodes high frequency features. Given the estimated Laplacians of the intermediate texture, the reconstruction process tries to obtain a new continuous surface texture which can reproduce the Laplacians. The reconstruction process uses the high-resolution grid previously generated for texture resampling.

A. A Weighted Laplacian Operator

The Laplacian of a vertex $\mathbf{v}_i$ in the high-resolution grid is computed by collecting the colors of its 1-ring neighbors as shown in Fig. 4. To compensate the non-uniform shape of the triangles, Fujiwara weights [20] are used:

$$L(\mathbf{v}_i) = -\sum_{0 \le j < N(i)} \frac{1}{e_{i_j}} (c_i - c_{i_j}), \qquad (1)$$

where $\mathbf{v}_{i_j}$ is a vertex directly connected to $\mathbf{v}_i$, e_{i_j} represents the edge length between $\mathbf{v}_i$ and $\mathbf{v}_{i_j}$, c_i and c_{i_j} represent the colors at the vertices, and $N(i)$ represents the number of

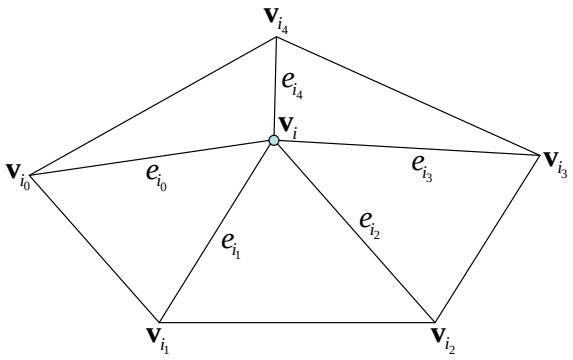


Fig. 4. The 1-ring structure of a vertex, $\mathbf{v}_i$.

neighboring vertices of $\mathbf{v}_i$. This Laplacian operator can also be rewritten as:

$$L(\mathbf{v}_i) = -\sum_{0 \le j < N(i)} L^f(\mathbf{v}_i, \mathbf{v}_{i_j}, \mathbf{v}_{i_{j+1}}), \quad (2)$$

where $L^f(\mathbf{v}_i, \mathbf{v}_{i_j}, \mathbf{v}_{i_{j+1}}) = \frac{1}{2e_{i_j}}(c_i - c_{i_j}) + \frac{1}{2e_{i_{j+1}}}(c_i - c_{i_{j+1}})$. $L^f(\mathbf{v}_i, \mathbf{v}_{i_j}, \mathbf{v}_{i_{j+1}})$ only involves the three vertices of a triangular face in the 1-ring structure of $\mathbf{v}_i$. This formulation allows us to consider the Laplacian operator as a summation of these facewise terms.

Since we would like to remove discontinuities along patch boundaries but still preserve original high frequency features within the texture patches, it is desirable to have spatially adaptive smoothing. To achieve this goal, we designed a weighted Laplacian operator which imposes potentially different weights on the edges. Eqs. (1) and (2) thus become

$$L_w(\mathbf{v}_i) = -\sum_{0 \le j < N(i)} \frac{w_{i_j}}{e_{i_j}}(c_i - c_{i_j}) \quad (3)$$

$$= -\sum_{0 \le j < N(i)} L_w^f(\mathbf{v}_i, \mathbf{v}_{i_j}, \mathbf{v}_{i_{j+1}}), \quad (4)$$

where w_{i_j} is a positive weight for the edge between $\mathbf{v}_i$ and $\mathbf{v}_{i_j}$, and $L_w^f(\mathbf{v}_i, \mathbf{v}_{i_j}, \mathbf{v}_{i_{j+1}})$ is also a facewise term similar to $L^f(\mathbf{v}_i, \mathbf{v}_{i_j}, \mathbf{v}_{i_{j+1}})$. If both $\mathbf{v}_i$ and $\mathbf{v}_{i_j}$ are from the same texture patch, we simply set $w_{i_j} = 1$; otherwise, w_{i_j} can be either smaller or larger than 1. If the weight of an edge is less than 1, the bonding between the two vertices of the edge is weakened, and there is less smoothing across the edge. Too small a weight may increase the stiffness of the resulting linear system discussed in the next section.

B. Laplacian Reconstruction

Given the Laplacians of the intermediate texture, we would like to reconstruct a new texture with the same Laplacians. Therefore, we set up a linear system with one equation per vertex. The equation for vertex $\mathbf{v}_i$ is expressed as

$$-\sum_{0 \le j < N(i)} \left(\frac{w_{i_j}}{2e_{i_j}}(c_i - c_{i_j}) + \frac{w_{i_{j+1}}}{2e_{i_{j+1}}}(c_i - c_{i_{j+1}})\right) = L_i, \quad (5)$$

where c_i and c_{i_j} represent unknown vertex colors in the new texture we would like to solve and L_i represents the estimated Laplacian of the intermediate texture at $\mathbf{v}_i$ using Eq. (3). The left hand side of this equation is actually the weighted Laplacian of the unknown new texture at $\mathbf{v}_i$. Since the weighted Laplacian is a linear operator, this equation is a linear equation of the unknown vertex colors. Note that if the textures have three color channels, there are three equations for each vertex. The collection of equations for all the vertices form a sparse linear system which has a symmetric coefficient matrix. Since the Laplacian operator is translation invariant, we need to fix the color of at least one vertex in order to obtain a unique solution of the linear system. Such fixed colors essentially form a boundary condition of the equations. Efficient iterative solvers [21] are a good choice for such a sparse linear system. In practice, we use a preconditioned (Incomplete Choleskey Factorization) conjugate gradient method [22].

1) Laplacian Estimation at Patch Boundaries: There are additional details concerning the estimation of the right hand side of Eq. (5) since the intermediate texture consists of patches with discontinuities on their boundaries. According to Eq. (3), the weighted Laplacian of a vertex can be estimated by accumulating a simpler term, $L_w^f(\mathbf{v}_i, \mathbf{v}_{i_j}, \mathbf{v}_{i_{j+1}})$, over all the triangular faces surrounding the vertex. However, a triangle may stride two or more texture patches. We summarize the estimation of this term as follows.

- If $\mathbf{v}_i$, $\mathbf{v}_{i_j}$ and $\mathbf{v}_{i_{j+1}}$ belong to the same patch, we directly use their colors to estimate $L_w^f(\mathbf{v}_i, \mathbf{v}_{i_j}, \mathbf{v}_{i_{j+1}})$.
- If the three vertices belong to two different patches, we should not directly use their existing colors because there may be a large gap among them. Since there must be a dominant patch having two of the three vertices, during the estimation of $L_w^f(\mathbf{v}_i, \mathbf{v}_{i_j}, \mathbf{v}_{i_{j+1}})$, the color of the third vertex should be taken from an extended version of the dominant patch to avoid large gaps.
- If the three vertices belong to three different patches, we simply randomly choose a dominant patch from the three. The colors of the other two vertices are taken from an extended version of that dominant patch.

2) Global Reconstruction: When the average brightness and color of the patches in the intermediate texture differ significantly (e.g. when they are from different complicated textures), we set up a sparse boundary condition and simultaneously solve the system of equations in (5) to remove the differences. As mentioned earlier, the boundary condition should consist of at least one constraint on the variables. A simple equality constraint is declared by setting the color of a vertex to be a fixed value. Such constraints reduce the number of variables in the linear system. The reduced linear system has a unique solution. The user can choose to interactively specify such constraints. In the absence of user-defined constraints, our program chooses to fix the colors at the centers of a random subset of the patches in the mesh. A more sophisticated constraint is defined by setting a linear combination of the vertex colors to be a fixed value. For example, we experimented with setting the average of all vertex colors to be a fixed value. Such linear constraints can be integrated into the linear system by considering them as additional equations. When there is exactly one linear constraint, the resulting enhanced linear system has a unique solution which defines a globally continuous new surface texture. When there is more than one

linear constraints, the system becomes overdetermined, and a least-squares solution should be obtained.

3) Local Reconstruction: When the average brightness and color of the patches in the intermediate texture are similar (e.g. when they are all from the same sample texture), locally reconstructing the texture can produce good results with much less computational cost than the global reconstruction. This is done by imposing color constraints on all the boundary vertices of the texture patches. The constrained color for every boundary vertex $\mathbf{v}_i$ is computed by simply blending the existing colors in its 1-ring structure,

$$c_{fixed}(\mathbf{v}_i) = \frac{1}{N(i)} \sum_{0 \le j < N(i)} c_{i_j}, \tag{6}$$

where the neighboring vertices of $\mathbf{v}_i$, $\{\mathbf{v}_{i_j}\}$, may belong to different patches. These dense color constraints effectively disconnect the texture patches from each other. The resulting linear system can be solved patch by patch. In practice, this scheme is a few times faster than the global reconstruction. Nevertheless, it only propagates information within each texture patch, which makes it better than local feathering along the patch boundaries but prevents it from resolving color differences on patches that are remote to each other. Therefore, this local scheme provides a tradeoff between quality and efficiency.

Fig. 5 demonstrates the visual quality of both local and global texture reconstruction. In the intermediate textures, there are obvious seams among the patches due to differences in low frequency components. Local Laplacian reconstruction can certainly remove these seams and create smooth transitions among the patches. However, it fails to produce large-scale changes that would make the base colors of the patches more consistent. On the other hand, global Laplacian reconstruction can perform such large-scale changes and produce more desirable results. To fully test the capability of global reconstruction, in the last example shown in Fig. 5, we artificially add a large random color shift to every texture patch in the intermediate texture. Global reconstruction can successfully remove these large color shifts and recover a consistent base color for all the patches. The reconstructed surface texture appears similar to the original input texture.

VI. Additional Experimental Results

We have conducted a large number of experiments on surface texture synthesis and mixing using the algorithm developed in this paper. Besides the examples shown in Fig. 5, we show a few additional results in Fig. 6. The first example in Fig. 6 gives a good demonstration on the fact that our algorithm can significantly improve the variability of the synthesized textures. The original FLOWERS texture has too few color variations. Extending such a texture over a large surface area would not make the result very appealing. By mixing it with the two leaf textures, the synthesized results become more interesting. In the first row of Fig. 6, the left one is the local reconstruction result while the right one is the global reconstruction result. In this particular case, both of them look interesting. The local result retains the rich colors of the three input textures while the global result has a smooth and subtle color change over the entire mesh.

TABLE I. The number of vertices and faces in the meshes and their corresponding fine grids used in our experiments.

	# mesh vertices/faces	# grid vertices/faces
Bunny	2503 / 5002	640258 / 1280512
Camel	2444 / 4884	625154 / 1250304
Pawn	510 / 1016	130050 / 260096
V-shape	170 / 336	43010 / 86016

TABLE II. The average running times (in seconds) of different stages of our algorithm on three meshes. These times were measured on a 3.2GHz AMD processor. "Initial" refers to the initial texture patch assignment stage; "Refinement" refers to the texture resampling and boundary refinement stage; "Local/Global Lapl" refers to local and global Laplacian texture reconstruction.

	Initial	Refinement	Local / Global Lapl
Bunny	8	5	13 / 25
Camel	3	4	13 / 28
Pawn	< 1	1	2 / 4.5

The statistics of the meshes used in our experiments are summarized in Table I. The running times of various stages of our algorithm are also summarized in Table II.

VII. Conclusions

In this paper, we proposed to decompose texture synthesis into two relatively disjoint stages. In the first stage, an intermediate synthesized texture is generated by only considering the high frequency details during neighborhood search and matching. In the second stage, we perform Laplacian texture reconstruction which retains the high frequency details but computes consistent low frequency components. It does not only affect texels close to discontinuities, but also modifies the rest of the texels. Therefore, it can be viewed as a global feature-preserving smoothing step, and is more effective than local feathering. Experiments indicate that our two-stage synthesis can produce desirable results for regular texture synthesis as well as texture mixing from multiple sources. In future, we would like to implement Laplacian reconstruction and other time-consuming steps on GPUs to achieve interactive performance.

References

[1] P. Perez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. on Graphics*, vol. 22, no. 3, pp. 313–318, 2003.

[2] Y. Yu, K. Zhou, D. Xu, X. Shi, H. Bao, B. Guo, and H.-Y. Shum, "Mesh editing with poisson-based gradient field manipulation," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 644–651, 2004.

[3] Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman, "Texture mixing and texture movie synthesis using statistical learning," *IEEE Transactions on Visualization and Computer Graphics*, vol. 7, no. 2, pp. 120–135, 2001.

[4] L. Liang, C. Liu, Y. Xu, B. Guo, and H.-Y. Shum, "Real-time texture synthesis using patch-based sampling," *ACM Trans. Graphics*, vol. 20, no. 3, pp. 127–150, 2001.

[5] A. Efros and W. Freeman, "Image quilting for texture synthesis and transfer," in *SIGGRAPH'01*, 2001, pp. 341–346.

[6] J. Dischler, K. Maritaud, B. Levy, and D. Ghazanfarpour, "Texture particles," *Computer Graphics Forum*, vol. 21, no. 3, pp. 401–410, 2002.

[7] V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick, "Graphcut textures: Image and video synthesis using graph cuts," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 277–286, 2003.

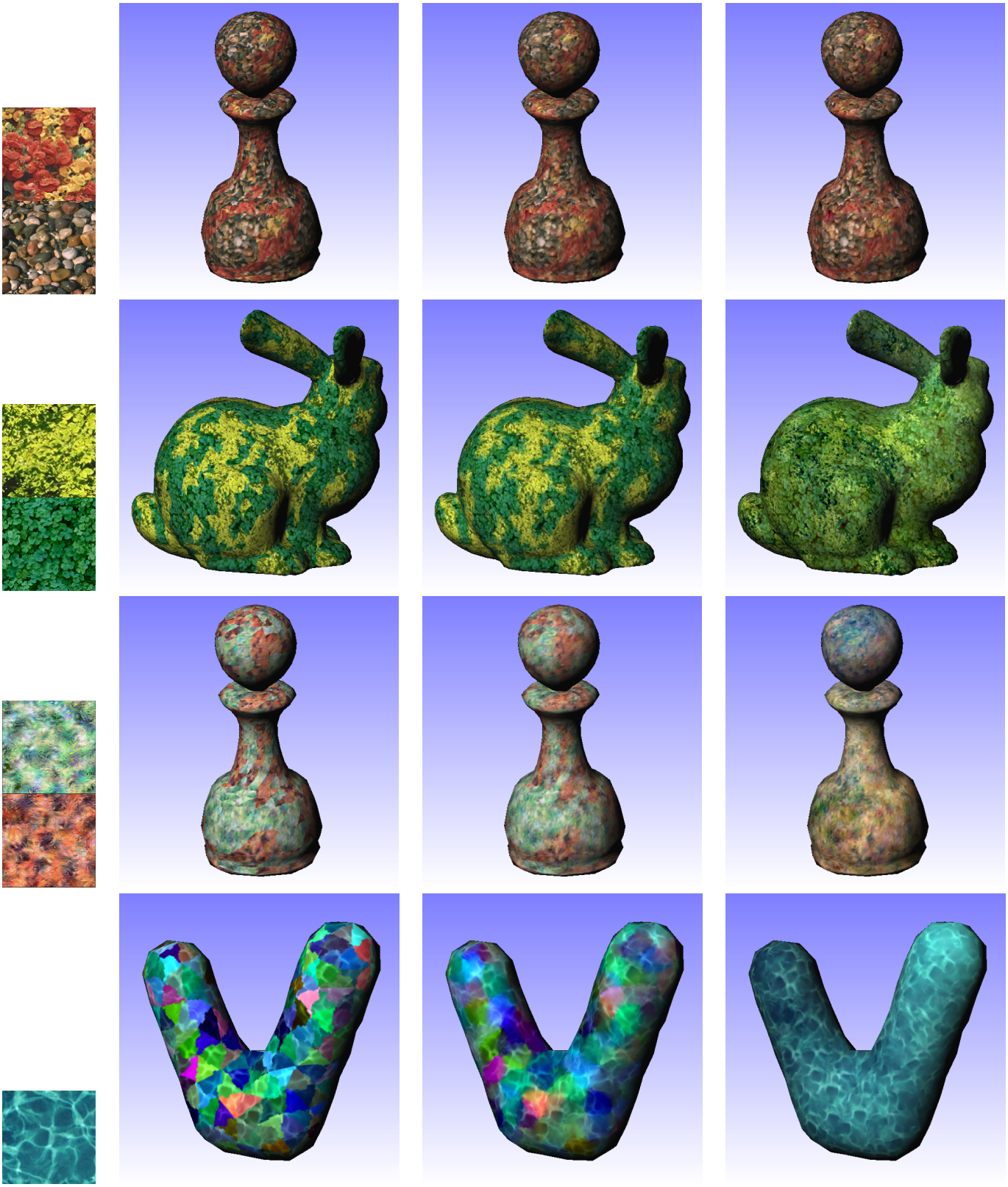


Fig. 5. (a) Input texture examples. (b) Synthesized intermediate textures with color discontinuities among patches. (c) Textures computed from local Laplacian reconstruction. (d) Textures computed from global Laplacian reconstruction. Note that local reconstruction works reasonably well for the texture mixture in the first row because the colors of the mixed texture patches are not too different. However, for the remaining three mixture examples, global reconstruction produces more natural and consistent low frequency components. The intermediate texture in the last row is artificially modified by adding a random color shift to each texture patch. Global texture reconstruction can successfully remove such color shifts.

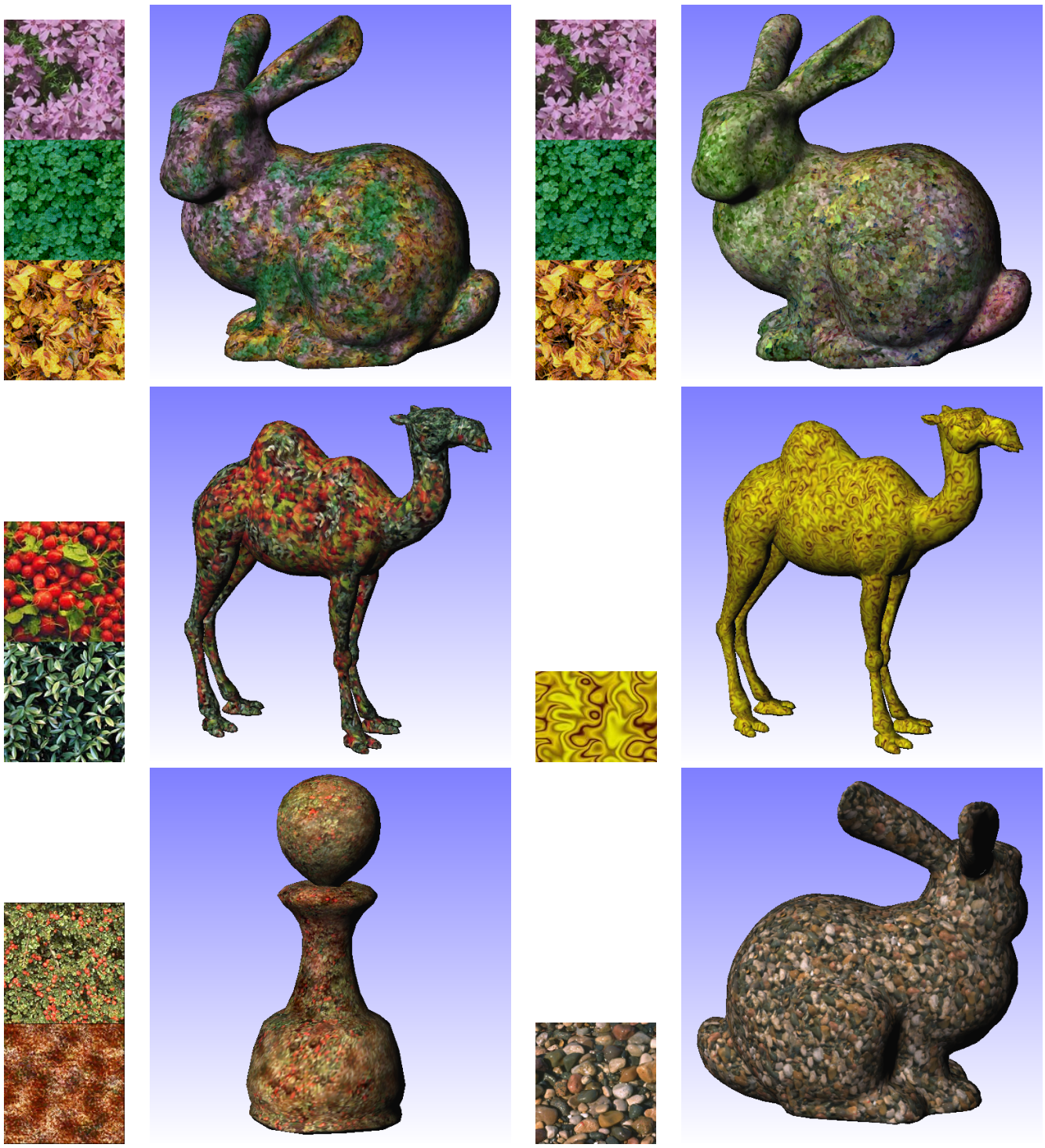


Fig. 6. Additional surface texture synthesis and mixing results.

[8] Q. Wu and Y. Yu, "Feature matching and deformation for texture synthesis," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 364–367, 2004.

[9] S. Lefebvre and H. Hoppe, "Parallel controllable texture synthesis," *ACM Transactions on Graphics*, vol. 24(3), pp. 777–786, 2005.

[10] C. Han, E. Risser, R. Ramamoorthi, and E. Grinspun, "Multiscale texture synthesis," *ACM Transactions on Graphics*, vol. 27(3), 2008.

[11] A. Rosenberger, D. Cohen-Or, and D. Lischinski, "Layered shape synthesis: automatic generation of control maps for non-stationary textures," in *ACM Transactions on Graphics (TOG)*, vol. 28, no. 5. ACM, 2009, p. 107.

[12] R. Wu, W. Wang, and Y. Yu, "Optimized synthesis of art patterns and layered textures," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 436–446, 2014.

[13] G. Turk, "Texture synthesis on surfaces," in *SIGGRAPH'01*, 2001, pp. 347–354.

[14] L.-Y. Wei and M. Levoy, "Texture synthesis over arbitrary manifold surfaces," in *SIGGRAPH'01*, 2001, pp. 355–360.

[15] C. Soler, M.-P. Cani, and A. Angelidis, "Hierarchical pattern mapping,"

in *SIGGRAPH'02*, 2002, pp. 673–680.

[16] J. Zhang, K. Zhou, L. Velho, B. Guo, and H.-Y. Shum, "Synthesis of progressively-variant textures on arbitrary surfaces," in *SIGGRAPH'03*, 2003, pp. 295–302.

[17] S. Magda and D. Kriegman, "Fast texture synthesis on arbitrary meshes," in *Eurographics Symposium on Rendering*, 2003, pp. 82–89.

[18] S. Zelinka and M. Garland, "Jump map-based interactive texture synthesis," *ACM Transactions on Graphics*, vol. 23, no. 4, pp. 929–1073, 2004.

[19] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. Intl. Conf. on Computer Vision*, 1998, pp. 836–846.

[20] K. Fujiwara, "Eigenvalues of laplacians on a closed riemannian manifold and its nets," in *Proceedings of the American Mathematical Society*, 1995, pp. 123:2585–2594.

[21] Y. Saad, *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, 1996.

[22] D. Kershaw, "The incomplete cholesky–conjugate gradient method for the iterative solution of systems of linear equations," *Journal of Computational Physics*, vol. 26, no. 1, pp. 43–65, 1978.