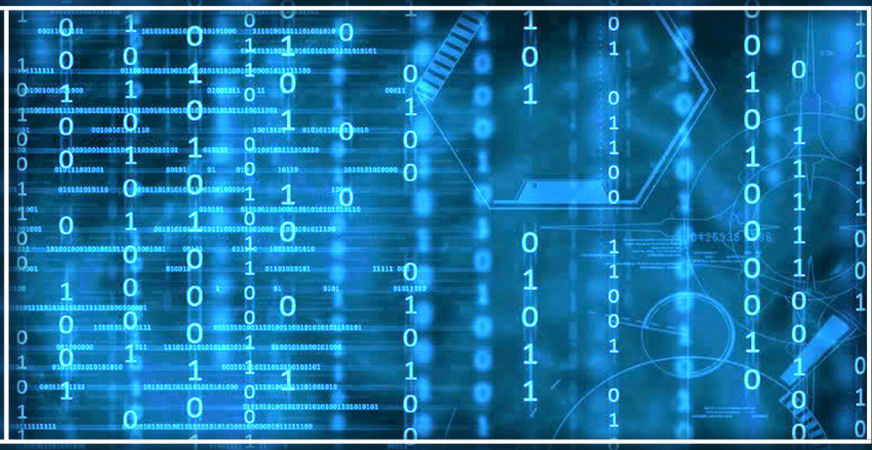
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 4 Issue 12

December 2013



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

SAI

www.ijacsa.thesai.org

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

Editorial Preface

From the Desk of Managing Editor…

It is our pleasure to present to you the December 2013 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 4 Issue 12 December 2013
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

- **A Kathirvel**
  Karpaga Vinayaka College of Engineering and Technology
- **A.V. Senthil Kumar**
  Hindusthan College of Arts and Science
- **Abbas Karimi**
  Islamic Azad University Arak Branch
- **Abdel-Hameed Badawy**
  Arkansas Tech University
- **Abdelghni Lakehal**
  Fsdm Sidi Mohammed Ben Abdellah University
- **Abdul Wahid**
  Gautam Buddha University
- **Abdul Hannan**
  Vivekanand College
- **Abdul Khader Jilani Saudagar**
  Al-Imam Muhammad Ibn Saud Islamic University
- **Abdur Rashid Khan**
  Gomal Unversity
- **Abeer Elkorny**
  Faculty of computers and information, Cairo University
- **Ahmed Boutejdar**
- **Dr. Ahmed Nabih Zaki Rashed**
  Menoufia University
- **Ajantha Herath**
  University of Fiji
- **Aderemi A. Atayero**
  Covenant University
- **Ahmed Mahmood**
- **Akbar Hossin**
- **Akram Belghith**
  University Of California, San Diego
- **Albert Alexander**
  Kongu Engineering College
- **Alcinia Zita Sampaio**
  Technical University of Lisbon
- **Ali Ismail Awad**
  Luleå University of Technology
- **Amit Verma**
  Department in Rayat & Bahra Engineering College
- **Amitava Biswas**
  Cisco Systems
- **Anand Nayyar**
  KCL Institute of Management and Technology, Jalandhar
- **Andi Wahju Rahardjo Emanuel**
  Maranatha Christian University, INDONESIA
- **Anirban Sarkar**
  National Institute of Technology, Durgapur, India
- **Andrews Samraj**
  Mahendra Engineering College
- **Arash Habibi Lashakri**
  University Technology Malaysia (UTM)
- **Aris Skander**
  Constantine University
- **Ashraf Mohammed Iqbal**
  Dalhousie University and Capital Health
- **Ashok Matani**
- **Ashraf Owis**
  Cairo University
- **Asoke Nath**
  St. Xaviers College
- **B R SARATH KUMAR**
  Lenora College of Engineering
- **Babatunde Opeoluwa Akinkunmi**
  University of Ibadan
- **Badre Bossoufi**
  University of Liege
- **Balakrushna Tripathy**
  VIT University
- **Basil Hamed**
  Islamic University of Gaza
- **Bharat Bhushan Agarwal**
  I.F.T.M.UNIVERSITY
- **Bharti Waman Gawali**
  Department of Computer Science & information
- **Bhanu Prasad Pinnamaneni**
  Rajalakshmi Engineering College; Matrix Vision GmbH
- **Bilian Song**
  LinkedIn
- **Brahim Raouyane**
  INPT
- **Brij Gupta**
  University of New Brunswick

- **ZHEFU SHI**
- **ZHIXIN CHEN**
  ILX Lightwave Corporation
- **ZLATKO STAPIC**
  University of Zagreb
- **ZUQING ZHU**
  University of Science and Technology of China
- **ZURAINI ISMAIL**
  Universiti Teknologi Malaysia

CONTENTS

Personalizing of Content Dissemination in Online Social Networks

Abeer ElKorany
Computer Science Department
Faculty of Computers & Information, Cairo University
5 Dr Ahmed Zoweil St., Orman, Giza, Egypt

Khaled ElBahnasy
Information Systems Department
Faculty of Computer & Information Sciences
Ain Shams University, Abbasia, Cairo,Egypt

***Abstract*— Online social networks have seen a rapid growth in recent years. A key aspect of many of such networks is that they are rich in content and social interactions. Users of social networks connect with each other and forming their own communities. With the evolution of huge communities hosted by such websites, users suffer from managing information overload and it is become hard to extract useful information. Thus, users need a mechanism to filter online social streams they receive as well as enable them to interact with most similar users. In this paper, we address the problem of personalizing dissemination of relevant information in knowledge sharing social network. The proposed framework identifies the most appropriate user(s) to receive specific post by calculating similarity between target user and others. Similarity between users within OSN is calculated based on users' social activity which is an integration of content published as well as social pattern Application of this framework to a representative subset of a large real-world social network: the user/community network of the blog service stack overflow is illustrated here. Experiments show that the proposed model outperform tradition similarity methods.**

Keywords—social network; content similarity measurement; Information retrieval; Information dissemination

I. INTRODUCTION

Nowadays, social networking has become an important part of online activities over the web. Social networks can be viewed as a structure which enables the dissemination of information through social interactions among individuals. The *analysis of the dynamics of such interaction* is a challenging problem in the field of social networks. Online social networks (OSN) such as Internet newsgroups, BBS, and chatrooms are interesting channels that enable its members to communicate and share activities in an easily accessible in anywhere any time trend. OSNs represent a new kind of information network that differs significantly from existing networks such as the Web. They are those network hosted by a web site where friendship represents shared interest or trust and online friends may have never met. When a user joins those networks, they could publish their own content, create links to other users in the network called "friends" or "acquainted". Virtual link is constructed between users with similar interests. User' generated information in OSNs has been characterized by either their provided published content in form of text, image or videos as well as network activities that are frequently changing over time. For example, web-blogging community is identified by its rich daily blog posts and a social network of bloggers who share, find, and disseminate content at a massive scale. Today a lot of active online social networks users complain that their streams have become too overloaded and hard to extract useful information from. To make use of the growing provided information flow within a social network and to keep being tuned with its related members, it is necessary to personalize the process of information distribution. Thus, it is necessary to control information propagation among users who share some common interest in OSN i.e finding similar users. Existing studies on user similarity focus on either link or content analysis. However, neither information alone is satisfactory in determining accurately the similarity between users. It is therefore important to unify the analysis of content and network activities about user in order to personalize content dissemination in social networks.

This research proposes a personalized user content dissemination framework that identifies the most appropriate user(s) to receive specific information (such as post) based on user similarity. Similarity between users within OSN is calculated based on users' social activity which is an integration of content published as well as social activities of users. Each of two sided of social pattern provides a partial indicator of the similarity. While content published by a user is used as an indicator of user interest, social pattern parameters tend to group people based on their similar networking behaviour. Accordingly, we identify, collect, and classify different users' online social activities that are used to construct their characteristics, and determine user' main preference. Next, users' preferences are used to detect similar users who are candidate to receive a specific stream (post). The vector space model is adapted to present user characteristics such that each user is represented by two vectors each contains a set of specific social activities of the user. The first vector represents the content published user through the *bag of word model* which is called content vector. Content vector represents the terms used in all the posts published by that user. While the second one, called social pattern vector, which represents the social pattern of user such as: user' contribution and influence attributes. The term weight of each vector is computed using different methods. For example, in content vector, all post published by each user is collected and TF-IDF is used to weight the terms. While, in the social pattern vector each social features is represented as term and weighted mean scheme is used to weight the terms. An aggregated linear model is then applied to combine similarity calculated by using each of those vectors. Thus, the process of content dissemination works as follow: first when a

target user post a stream, cosine similarity is applied to compute similarity between that post and content vector of all users and only the top ranked 20 users are used as input to the next phase. Next, social pattern vector of each of those 20 users is retrieved and cosine similarity is used again to compute similarity between social pattern vector of target user and the top 20 users in order to re-rank them. It is significant to mention that the proposed process is generally applicable to any knowledge sharing environment. However, our work is supported by a set of experiments and tests conducted. The experiment is applied on real world weblog system (question/answering) Stack Overflow dataset [1] a question answering web community that allows users to ask and answer questions about computer programming languages. A modified 5-fold cross validation method is used to insure the accuracy of the proposed model. This paper is organized as follow: section2 presents related works in three main areas such as social networking services, information filtering, and other methods used for user similarity. Section3: discuss the proposed model for personalization content dissemination and section4 explains the main components of the system used to identify candidate users to receive a specific post. Section5 discuss the experiments hold to measure the accuracy and efficiency of the proposed model on real dataset and finally section6 conclude the work and propose future work.

II. Related Work

Our model is closely related to models of information filtering and data analysis in social network

A. Social network service

Several Social Networking Services (SNS) are designed to facilitate communication, collaboration, and content sharing over a network of contacts [1]. They enable users to share profiles and personal information, media, event planning, communicate by email, send instant messages, share announcements, blog together, creation of interest groups, and meet online with their friends or even other new people. In SNS, the variety and quality of content is a key factor for success. Encouragement to produce content is therefore commonly observed in these networks. Therefore, analysis of two primary kinds of data in the context of social networks is widely increased. These data are:

Linkage-based and Structural Analysis: In linkage-based and structural analysis, analysis of the linkage behavior of the network is applied in order to determine important nodes, communities, links, and evolving regions of the network. Such analysis provides a good overview of the global evolution behaviour of the underlying network.

Adding Content-based Analysis: Many social networks such as *Flickr*, *Message Networks*, and *Youtube* contain a tremendous amount of content which can be leveraged in order to improve the quality of the analysis. For example, a photograph sharing site such as *Flickr* contains a tremendous amount of text and image information in the form of user-tags and images. Similarly, blog networks, email networks and message boards contain text content which are linked to one another.

B. Information filtering in social network

Information overloading has been a major problem in social media. Thus, social filtering systems are used to filter social media streams in order to overcome this problem . Several information filtering systems have been proposed such as in [2] which utilize text of documents that the user is interest in with other sources of information to identify social features of the users. These features are then used to detect others who are more likely to post relevant content. However, limitation on text size in micro-blogging services result in sparseness of data for text classification. Another systems suggested to use Wikipedia as an external source [3,4], or using web search engine results[5] to enhance text classification. Another approach that used topic modelling techniques [6] such as Latent Dirichlet Allocation [7] to classify texts based on universal corpus . Another work focused on short texts is [8], where a method for measuring the semantic similarity of texts, using corpus-based and knowledge-based measures of similarity is proposed.

C. User Similarity in social networks

There have been numerous efforts to calculate the user similarity for different objectives such as recommending people. Guy et al. [9] proposed a method based on various aggregated information about people relationships and focused only on people that the user is already familiar with. Thus, this method was not used for calculating the similarity with an unknown user such as suggest a new friend in the online social network. Terveen et al. [10] proposed a framework called *socialmatching* that match people mainly using their physical locations. Other system focused on applying the semantics of the location in order to calculate similarity between users by capturing the user's intention and interest and considering the similarity between different locations using the hierarchical location category[11]Other methods utilize similarity measurement in social network to recommend experts. McDonald et al. [12] proposed an expert locating system that recommends people for possible collaboration within a work place. An expert search engine was described in [13] which found relevance people according to query keywords. Those approaches are useful to find co-workers or experts in a specific domain however; they cannot be used for finding similar users in general.

III. Personalized Flow of Information in Online Social Network

OSN have exploded in popularity. They could be classified as into two categories, the networking oriented OSNs and knowledge-sharing oriented OSNs. The former, such as Facebook and LinkedIn, emphasizes more on the networking perspective, and the social relationship is the basis of these OSNs. Hence, they are called networking oriented OSNs. While the latter, such as blog networks, question answering networks, and viral video networks, emphasizes more on the knowledge or content sharing [14]. In knowledge-sharing OSNs, issues such as users' participation in network and their generated content are crucial to healthy growth of those networks. Information overloading is a major problem in such

[1] Stack overflow. http://stackoverflow.com

knowledge sharing environments [15]. Thus, the proposed framework shown in figure1 aims to overcome the information overloading by enhancing the distribution of content among users. It suggests the most relevant users that are candidate to receive a specific post from a target user. By identifying main features that characterize users and determining users preference, personalization of flow of information is achieved which is illustrated in this section.

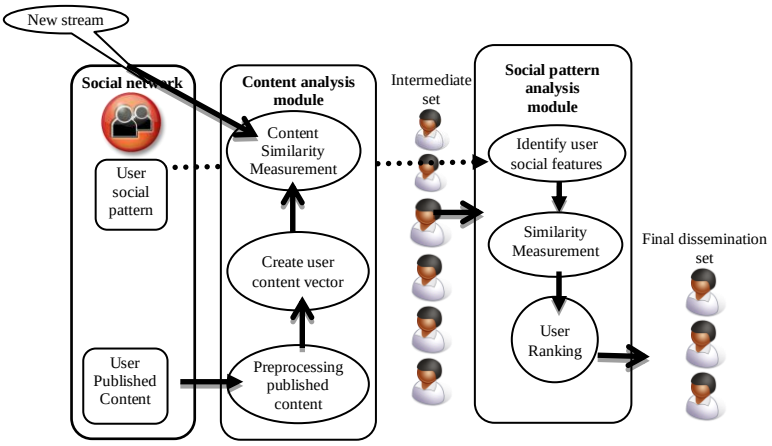


Fig. 1. Framework of Personalized Information flow

A. *Analysis of users' social contributions*

User contribution with a social network is changing over time in form of interaction and content provision resulting in variant network structure and text dissemination which evolve simultaneously and interrelated. User's activities in OSNs include authoring content, viewing, and networking. It has been also assumed that there is strong correlation between user active time and user contribution [16,17]. In general, user within the same OSNs could be classified as online and offline users. Online users are those who always active in writing posts, comments, view others publish content, provide feedback in form of like or dislike and many other activities. While offline users are those who just visit their homepages on daily or weekly basis without any contribution. Most of existing approaches for user and content recommendation rely only on network structure and relationship between users without considering the published content. Our proposed framework identifies user's preference in OSN by considering two main folds: content published by users and user social pattern. In blogs and question answering networks, there are two important elements in a shared content: text and hyper-links which provide related information from external sources such as web pages or images. Similar users are characterized by posting similar text and hyper-links. Therefore, posts of each user are parsed, and a bag of words is created for each user, keyword index is computed indicating that the more terms two users share, the stronger the tie between them. Second, other network features that distinguish users in OSN are considered which we classify as : user contribution and user influence are identified in order to impose a finer grained similarity between users. Each of them is expressed by a set of weighted features extracted from social activities of users which will be explained in details in next section. Unlike other approaches that only consider user pre-defined attributes such as demographical attributes, geographic location, and defined interests, our approach relies on extracting social activities and calculate their weight according to user contribution.

B. *Multilevel Model for detecting Relevance Users*

Recently, OSN started to be modelled with *rich structured data* that incorporate *semantics*. In such models edges between users are split to weighted links based several features such as: communication aspects (uni-direction or multi-direction), frequency of communication, and influence measure which are used to build up clusters of similar user. Therefore, we calculate similarity between users through aggregating communicative content of users in form of mutual published content with social activities. Our model of detecting relevance users is based on the assumption that if users *X* and *Y have n* similar published contents and similar social features, they have strong tie. In the proposed framework, adaptive vector space model is used to compute social similarity measurement of users. The vector space model is a standard and effective algebraic model widely used in information retrieval (IR) that use Cosine similarity to compute relevancy of documents with respect to a given query. Accordingly, our model works by first identifying content and social features, collecting required information, creating corresponding vectors. Each user is expressed by two vectors of elements corresponding to their published contents and social pattern respectively. A two-level model is then used to identify most candidate users to receive a specific post. Each level is responsible on computing similarity between users using different user social features (stored in different vectors) and a linear model is then used to combine similarity generated from each level. The main idea is based on utilizing model-based collaborative filtering by first finding users with similar content and then utilizes social features to map social characteristics of users and get the most neighbourhood users for the target user.

IV. Architecture of Personalized Information Diffusion

The main objective of the proposed model is to improve the dissemination of information in knowledge sharing network by identifying the most appropriate (similar) users to this information as well as its publisher. In social media, users are identified through their social content and participation. Therefore, the proposed personalization framework is decomposed of two main modules: content similarity measure and social patterns similarity measure.

1) User content similarity: using content of the posts and comments generated by users, similarity measurement is applied to get top ranked 20 users with respect to specific post. This short list of users represents the closest users to that stream.

2) User social pattern similarity: additional social information about top ranked 20 users is used to ensure similarity "coverage". Cosine similarity is then applied between social pattern vector of target user (who post that

stream) and the those users to rank the most neighbourhood users

Before content similarity phase takes place, pre-processing of the content representing the posts of users is applied as shown in figure1.

B. Pre-processing

Currently, there is a large volume of text data that is produced from the social communities such as blogs, tweets, and comments. Therefore, during this step, all blog- posts are aggregated from all users to form a corpus. Posts are parsed in order to clear all special characters, numbers, dates, stop words and single characters. This yields to construct a vocabulary that represents the set of words that have been used by the whole users of the social network within a specific time period. The final step of the preprocessing decomposes stemming, identifies hyper-links or code which may be included in a user posts, and then constructs content vector of each user.

C. User content similarity measurement

In linear algebra a *vector space* is a set *V* of *vectors* together with the operations of addition and scalar multiplication. A vector space model (VSM) is an algebraic model introduced a long time ago by Salton [18] in the information retrieval (IR) field. In a more general sense, a VSM allows to describe and compare objects using N-dimensional vectors. Each dimension corresponds to an orthogonal feature of the object (e.g. traditionally weight of certain term in a document). We adapt the vector space model so that it treats each user as a document and her posts as terms disregarding grammar and even word order. In IR field there are several approaches to calculate the weight of a term in a document. In our model, we apply the *term frequency-inverse document frequency (tf-idf).* Term Frequency (tf) assigns the weight to be equal to the number of occurrences of the term *t* in document *d*. While *IDFt* is obtained by dividing *N* by *DFt* and then taking the logarithm of that quotient, where *N* is the total number of posts generated by a user and *DFt* is the post frequency of *t*, i.e., the number of posts containing the term *t*. This approach identify the rarity of *t* in a given corpus Thus, if *t* is rare, then the posts containing *t* are more relevant to *t* which match with the idea we propose to find the target (most similar users) to specific set of words(post). Thus, content vector of a user would represent all words of her/his posts associated with TF-IDF weights ***wjt*** , where *wjt* is the weight of word *t* in post *j* and is calculated as follow:

$$wjt = tf(t, j) \log(\text{n posts}/df(t))$$

Then, when a user post a specific query (post) Cosine similarity is used as a standard measure estimating relevancy between this post and other users' content vector. The top 20 similar users to that post are selected to be used in the next phase.

D. Social patterns similarity measurement

In social knowledge sharing, in order to identify similar users, it is significant to consider the effect of network relationships. By aggregating communicative activities of users in form of social interaction, hidden relations between users are discovered, and hence a list of most similar users is generated. Unlike other approaches which considers only number of mutual friends [19], our proposed social pattern similarity measures consider all user social activities which we classify into two categories. The first category covers user contribution with the system while the second covers user influence with other community's members. User contribution is measured by the frequency of contribution to the knowledge sharing network (in our case blogs) which is: average number of posts and average number of comments a user provides. On the other hand, bloggers tend to *interact* with other bloggers by providing comment, like, or favorite in response to specific blog posts. Thus, we consider this type of information as a measure of user influence or trust relationship. "Trust relationships" are different from "social friendships" in many aspects. For example, when a user u_i likes a blog issued by another user u_j, user u_i probably will add user u_j to his/her trust list. The study of homophily has shown that people with similar interests are more likely to become connected, associate, and bond with other similar users [20]. Based on that hypothesis, we measure the user influence by the attributes that reflect the recognition he got from others in the same social network which may vary from social network to another. For example: Endorsements are a one-click system to recognize someone for their skills and expertise on LinkedIn, the largest professional online social network. In our case of stackoverflow, we map existing social features with the idea of "Trust relationships". Therefore, we utilize the following attribute:

`ViewCount`: Number of views of posts the users obtained
`FavoriteCount`: Number of users, who set a user's posts as favorite,
Vote: Count the number of votes for specific user' posts,

Each user is represents as vector associated with scores representing the average weight of her/his social features as terms. This was accomplished through combining all previous posts and comments of each user and calculated the average number of views, score, and favorite she obtained from others as well as the average number of posts and comments published by that user. Two users are similar if their vectors differs only a few coordinates. A high degree for a preference (term) of a user can be interpreted in a way that the other user repeatedly (frequently) confirms her preference [21].

V. EXPERIMENT

In this section we present several experiments to show how the proposed similarity measures model affect the dissemination of information in online social network. As mentioned earlier, the model aims to generated the minimum and appropriate users who can receive a specific information based on content-matching with that post and social pattern matching with the seed user who posted it. We applied this model on data dump provided by Stack Overflow[2] which is an online platform where users can exchange knowledge related to programming and software engineering tasks. This platform combines features of Wikis, Blogs and Forums, and aims to

[1] Stack overflow data dump http://blog.stackoverflow.com/2009/06/stack-overflow-creative-commons-data-dump/. /. Accessed January 2013

provide free knowledge sharing between software developers worldwide. The collected data contains all the questions and answers posted on the web site between July 31, 2008 and March 31, 2009. All posts, both questions and answers, are scored, viewed, and voted by the users. Some questions are set as favorites by some users. We use a subset of the posts collection, obtained by filtering the document collection to select the 100,000 posts belongs to 2000 different users. Several experiments are applied, thus we split the set of users into 5 equally sized disjoint groups of users from G1 to G5 each of 400 users {Gi=1-5}. Two main experiments have been applied, the first one aims to prove the accuracy of the proposed model which the second one target the efficiency.

A. *Experiment Setup*

The first experiment has been applied using content-similarity module and trying different weighting scheme using the vector space mode. We used the TF and TF/IDF to weight the terms and according to table1, the similarity values obtained when using TF/IDF is better for top 20 users.

TABLE I. COMPARISON BETWEEN DIFFERENT CONTENT WEIGHTING SCHEME

	TF/IDF	TF		TF/IDF	TF
User1	0.6144	0.5987	User11	0.5947	0.2537
User2	0.6082	0.5973	User12	0.5942	0.2339
User3	0.6071	0.5879	User13	0.5928	0.2303
User4	0.6021	0.4740	User14	0.5923	0.2273
User5	0.6020	0.3762	User15	0.5912	0.2239
User6	0.5999	0.3605	User16	0.5910	0.2171
User7	0.5989	0.3538	User17	0.5902	0.2167
User8	0.5988	0.3316	User18	0.5898	0.1994
User9	0.5986	0.3180	User19	0.5897	0.1983
User10	0.5969	0.3148	User20	0.5893	0.1898

In order to ensure the accuracy of the proposed system, we adapt the cross validation. Thus, we create a training set and sets. The training set contains one group of users (G1=400 users) and other test sets are created based on different combinations between G1 and other groups. We use mathematical **combination** to produce several test sets. This combinations imply that if the set has *n* elements the number of *k*-combinations is equal to the binomial coefficient which can be written using factorials as:

$$\text{Number of generated groups} = \sum_{k=1}^{k=5} \frac{n!}{k!(n-k)} \qquad \text{Equ(1)}$$

As per equation1 and when having N=5, 31 subsets are generated. However, only 16 of them contain target group G1 like as follows:

$\{G_{1}\}$,
$\{G_1 \cup G_2\}$, $\{G_1 \cup G_3\}$, $\{G_1 \cup G_4\}$, $\{G_1 \cup G_5\}$,
$\{G_1 \cup G_2 \cup G_3\}$, $\{G_1 \cup G_2 \cup G_4\}$,$\{G_1 \cup G_2 \cup G_5\}$, $\{G_1 \cup G_3 \cup G_4\}$, $\{G_1 \cup G_3 \cup G_5\}$, $\{G_1 \cup G_4 \cup G_5\}$
$\{G_1 \cup G_2 \cup G_3 \cup G_4\}$, $\{G_1 \cup G_2 \cup G_3 \cup G_5\}$, $\{G_2 \cup G_3 \cup G_4 \cup G_5\}$,$\{G_1 \cup G_3 \cup G_4 \cup G_5\}$,
$\{G1 \cup G2 \cup G3 \cup G4 \cup G5\}\}$.

Next, we select a random post generated by a random user and apply content-similarity phase to get the top ranked 20 users relevant to that post as shown in first column in table2 using TF/IDF method. Then, we regularly increase the neighbourhood of a target $user_x$ and find out whether the system would produce the same set of candidate users and check the similarity values they obtained. Therefore, we use the same post for the same user all over other 15 groups and get similarity value for the same set of users. According to table2, similarity value obtained for those users remain almost the same however the number of neighbour is. This matches our assumption regarding distributing of information which should target interest of users however the size of community. Most relevant users to a specific stream (post) are filtered and posts propagate to specific users and thus we can overcome the problem of information overloading.

Next, in order to be able to detect whether the proposed similarity approach is able to reveal the most candidate users based on the content features, we apply the Mean Average Error (MAE). MAE is used here to measure the average absolute deviation between a predicted set of users among several neighborhood users.

Therefore, we repeat the previous experiment and get top 10 users for each of the 16 groups and get their similarity values. As shown in figure2 which represents the MEA between content similarities for top 10 users in all groups. According to the figure, the difference between similarity values is minor however the group size is which ensure the accuracy of user prediction.

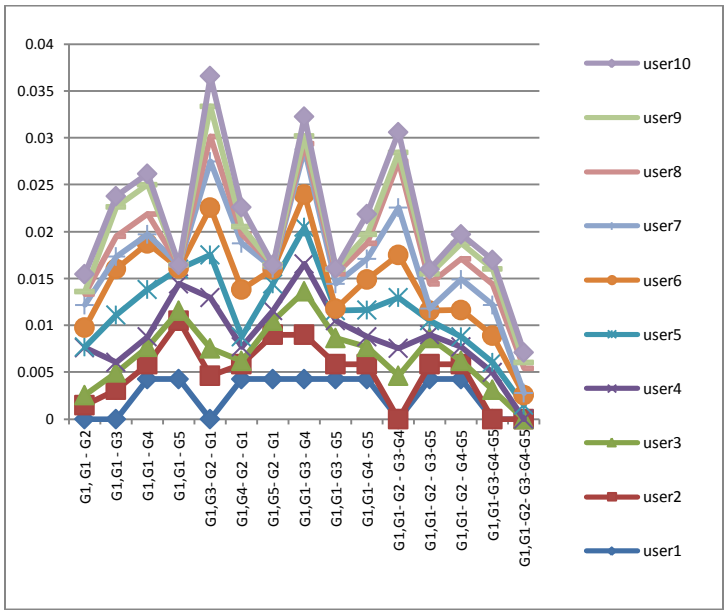


Fig. 2. MEA of content similarities (Top 10 users)

B. *Experimental evaluation criteria*

Other experiments have been applied to measure the efficiency of the proposed model.

Therefore, we first apply the content similarity module, get top 20 users similar to a random post, get their similarity values, and finally apply the social similarity module and get top 10 users. This experiment shows the effect of using addition social pattern features.

TABLE III. User Similarity (Top 10)

Model -rank	model-similarity value	Content-similarity -rank	content-similarity -value
1	0.99996	19	0.60820
2	0.99989	1	0.58928
3	0.99989	15	0.59993
4	0.99987	4	0.59016
5	0.99982	7	0.59225
6	0.99980	6	0.59118
7	0.99975	9	0.59423
8	0.99968	12	0.59860
9	0.99961	14	0.59886
10	0.99959	20	0.61443

As per table3, similarity between users and target user$_x$ who post the stream significantly improved when applying the whole model. Furthermore, ranking of top users has been changed when using the additional features which reflect the fact that we should consider network features of users when measuring similarity. For example, after adding network data user1 was ranked 19 using content similarity only while user 2 was having the first place Next, in order to measure the accuracy of the whole model, we repeat the first experiment l by applying the whole model and obtain similarity value of top 10 users. According to Figure3, similarity value has been improved after applying the social pattern level. Thus, we get the top candidate users generated from content-similarity module from table2, get their social similarity values and filter the top 10 users for all test groups. It is significant to mention that having applied the social pattern similarity module outperform using only content similarity.

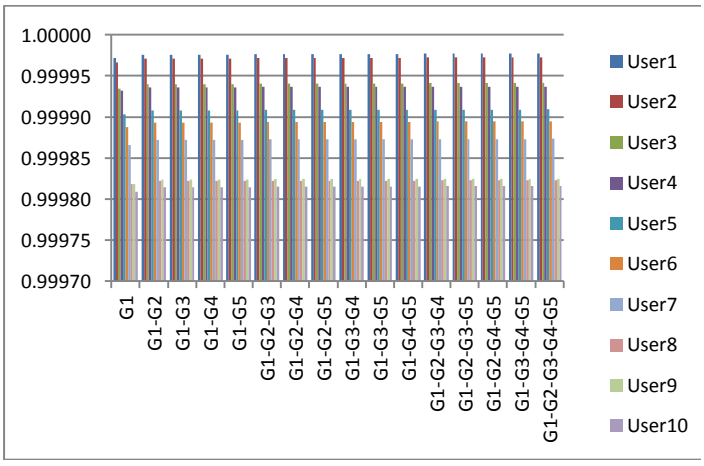


Fig. 3. User Social pattern similarity (Top 10)

Furthermore, longest common subsequence [LCS] method is used to measure the ranking used in our proposed model. Thus, we get the sequence of the top 10 users obtained from group1 containing 400 users, we apply the model all over other 15 groups and get the order of those users in each group as shown in figure4. According to figure4, there is no intersection points in the plotted area which mean that the order of predicted users remains the same however the size of the neighbourhood is.

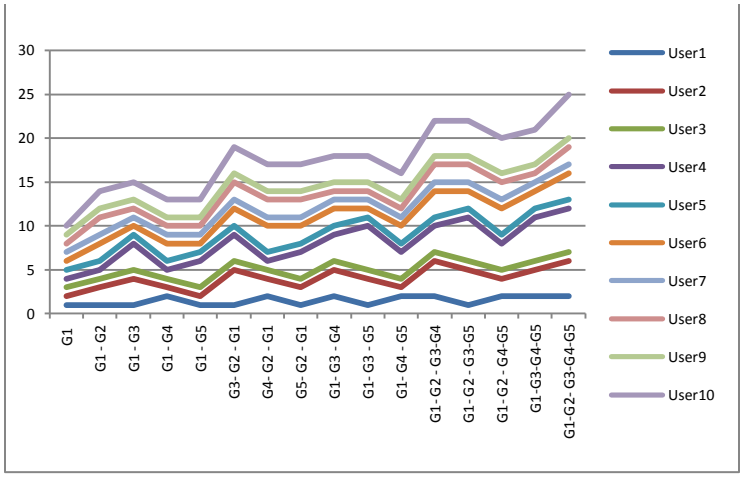


Fig. 4. longest common subsequence between users over groups (Top 10)

VI. CONCLUSION

This research proposes a model for enhancing the personalization of content dissemination among users in online social network with the aim of overcome the information overloading problem. The proposed similarity measurement model utilizes users' characteristics in social network. Such that when a target user posts a text, a set of most candidate receivers is generated and ranked by considering both similarity between this post and their interest as well as relevancy between social pattern of target user and others. Interest of users is extracted from content published by a user while social pattern is identified based on her social activities. Each user is represented by two vectors correspond to content and social pattern activities respectively. Different term weighting scheme was applied in order to weight social features and cosine similarity is then used to order the most similar users that is candidate to obtain that stream and to communicate with. The proposed model is applied on real dataset from stackoverflow and the experimental show that the accuracy of proposed method is precise as we obtain almost the same set of candidate users however the size of neighbourhood is. Furthermore, adding social pattern features in measuring similarity among user increase the similarity scores. In order to enhance the model, ontology could be used to measure content similarity among users. Furthermore, other topological features could also be used to rank users.

REFERENCE

[1] Cachia, R. (2008). Social Computing: The Case of Online Social Networking. IPTS Exploratory Research on Social Computing. JRC Scientific and Technical Reports.

[2] Güç, B. (2010). Information filtering on micro-blogging services (Doctoral dissertation, Swiss Federal Institute of Technology Zurich, Institute of Information Systems).

[3] Banerjee, S., Ramanathan, K., & Gupta, A. (2007, July). Clustering short texts using wikipedia. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 787-788). ACM.

[4] Schönhofen, P. (2009). Identifying document topics using the Wikipedia category network. Web Intelligence and Agent Systems, 7(2), 195-207.

[5] Bollegala, D., Matsuo, Y., & Ishizuka, M. (2007, May). Measuring semantic similarity between words using web search engines. In Proceedings of WWW (Vol. 766).

[6] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In WWW '08: Proceeding of the 17th international conference on World Wide Web, pages 91–100, New York, NY, USA, 2008.

[7] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

[8] Mihalcea, R., Corley, C., & Strapparava, C. (2006, July). Corpus-based and knowledge-based measures of text semantic similarity. In Proceedings of the national conference on artificial intelligence (Vol. 21, No. 1, p. 775). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

[9] Guy, I., Ronen, I., Wilcox, E.: Do you know?: recommending people to invite into your social network. In: International Conference on Intelligent User Interfaces, pp. 77–86 (2009) Terveen, L.G., McDonald, D.W.: Social matching: A framework and research agenda. ACM Trans. Comput. -Hum. Interact, 401–434 (2005)

[10] Lee, M. J., & Chung, C. W. (2011, January). A user similarity calculation based on the location for social network services. In Database Systems for Advanced Applications (pp. 38-52). Springer Berlin Heidelberg

[11] McDonald, D.W, 2003, Recommending collaboration with social networks: a comparative evaluation. In: Conference on Human Factors in Computing Systems, pp. 593–600 (2003)

[12] Ehrlich, K., Lin, C.Y., Griffiths-Fisher, V.: Searching for experts in the enterprise: combining text and social network analysis. In: International ACM SIGGROUP Conference on Supporting Group Work, pp. 117–126 (2007)

[13] Guo, L., Tan, E., Chen, S., Zhang, X., & Zhao, Y. E. (2009, June). Analyzing patterns of user content generation in online social networks. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 369-378). ACM.

[14] Adomavicius, G. and Tuzhilin, A. 2005. Personalization technologies: a process-oriented perspective. ACM. 48, 10, 83-90.

Guo, L., S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang. Measurements, analysis, and modeling of BitTorrent-like systems. In Proc. of ACM SIGCOMM IMC, 2005

[15] Leskovec, J L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In Proc. Of ACM SIGKDD, 2008

[16] Salton, G The smart retrieval system. Experiments in Automatic Document Processing, 1971

[17] Akcora, C. G., Carminati, B., & Ferrari, E. (2013). User similarities on social networks. Social Network Analysis and Mining, 1-21.

[18] Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social networks*, *25*(3), 211-230

[19] Wang, Shuaiqiang and Sun, Jiankai and Gao, Byron J and Ma, Jun}, Adapting vector space model to ranking-based collaborative filtering}, Proceedings of the 21st ACM international conference on Information and knowledge management}, pp{1487--1491}, 2012

TABLE II. USER CONTENT SIMILARITY (TOP 20)

	G1	G1 - G2	G1 - G3	G1 - G4	G1 - G5	G1 - G2 - G3	G1 - G2 - G4	G1 - G2 - G5	G1 - G3 - G4	G1 - G3- G5	G1 - G4 - G5	G1 - G2 - G3-G4	G1 - G2 - G3-G5	G1 - G3 - G4-G5	G1, G2,G4,G5	G1,G2 G3,G4,G5
user1	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144	0.6144
user2	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082	0.6082
user3	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071	0.6071
user4	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021	0.6021
user5	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020	0.6020
user6	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999	0.5999
user7	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989	0.5989
user8	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988	0.5988
user9	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986
user10	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969	0.5969
user11	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947	0.5947
user12	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942	0.5942
user13	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928	0.5928
user14	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923	0.5923
user15	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912	0.5912
user16	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910	0.5910
user17	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902	0.5902
user18	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898	0.5898
user19	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897	0.5897
user20	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893	0.5893

Attacking Misaligned Power Tracks Using Fourth-Order Cumulant

Eng. Mustafa M. Shiple
Electronics Department
National Telecommunication Institute Cairo, Egypt

Prof. Dr. Iman S. Ashour
Electronics Department
National Telecommunication Institute Cairo, Egypt

Prof. Dr. Abdelhady A. Ammar,
Communication Department
Al Azhar University
Cairo, Egypt

***Abstract*—Side channel attacks (SCA) use the leaked confidential data to reveal the cipher key. Power consumptions, electromagnetic emissions, and operation timing of cryptographic hardware are examples of measurable parameters (analysis) effected by internal confident data. To prevent such attacks, SCA countermeasures are implemented. Misaligned power tracks is a considerable countermeasure which directly affect the effectiveness of SCA. Added to that, SCA are suffering from tremendous types of noise problems. This paper proposes Fourth-order Cumulant Analysis as preprocessing step to align power tracks dynamically and partially. Moreover, this paper illustrates that the proposed analysis can efficiently deal with Gaussian noise and misaligned tracks through comprehensive analysis of an AES 128 bit block cipher.**

Keywords—Correlation power analysis (CPA); Differential power analysis (DPA); side channel attack; FPGA; AES; cryptography; cipher; fourth-order cumulant; Gaussian noise; higher order statistics

I. INTRODUCTION

In the past decade, new threats become more and more efficient and powerful against cryptosystems. There are three categories of attacks, active invasive (e.g. fault injection), active non-invasive (e.g. tampering), and passive (power consumption, timing attack) [1].

Invasive attack bases on penetration the Device Under Attack (DUA) package and analyzes each layer to modify or monitoring the entire signals and buses. The countermeasures of this type is based on burying critical layers beneath other layers of conducting metal layers to avoid direct connection from the surface. Added to that, distributing sensors all over the chip to detect any attack trials. These sensors erase the memories and all critical registers when activated. Ffurthermore, the designers use the nonstandard cells, scrambling the bus implementation, scrambling the stored data, dummy structure to mislead the attackers to discover the design architecture.

Like invasive attacks, Semi-invasive attack requires depackaging the chip without creating contacts to the internal lines [2]. Ultraviolet ray is used to unauthorized access to the stored data meanwhile the cryptosystem designers use many of anti-fuse to prevent such attacks like Security Fuse, Program Fuse, Array Fuses, and oProbe Fuse [3]. Another technique; a transient fault during the execution of some process is injected. A fault allows bypassing security condition checks, such as PIN correctness or in some other cases reducing the cipher rounds by manipulating the cipher counter. Many researchers have mounted differential fault analysis (DFA) on symmetric key encryption algorithms, such as the triple-DES [4], Advanced Encryption Standard (AES)[5], CLEFIA [6], and ARIA [7].

On the other hand, Noninvasive attack does not depackaging the DUA since the main load of this attack derives towards exploiting the confidential data through observing the output behavior of the DUA. Noninvasive attack could be Passive, also called side-channel attacks, or Active attack. Side-channel attack does not involve any interaction with the attacked device but, usually, observation of its power consumptions and electromagnetic emissions.

Paul Kocher et al. introduced a powerful cryptanalysis technique called Differential power analysis (DPA) in 1999 [8]. This technique is based on the dependency of the processed data/the operation performed to power consumption of the device under attack (DUA). Kocher proofed that the leak information could easily reveal the confidential cipher key.

Many different countermeasures are emerged to stop SCA and secure the cryptosystems. Misaligned tracks, adding non-correlated noise, and breaking the relation between processed data and measured parameters are designers targets to protect their hardware.

A novel approach is proposed to dynamically align power traces and reducing noise signal effects in one shot. Moreover, the proposed idea shows a great improvement in processing time reaches more than 75% less than other techniques. By attacking a FPGA-based 128 bit AES block cipher, Analysis results show that the proposed idea efficiently helps SCA in harsh environment (noisy and suffering from alignment problems).

This paper is organized as follows. An overview on CPA is provided in Section II. In Section III, IV, Side channel attack Noise and Alignment techniques are presented. Then, in Section V, The background of higher order statistics is briefly introduced. Section VI provides a detailed explanation of the proposed method. Finally, we conclude with the main advantages presented in this paper.

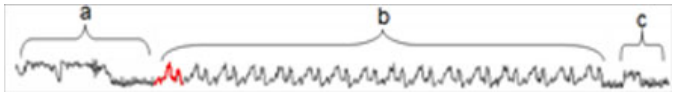


Fig. 1. SPA for DES algorithm; (a) 16 Rounds, (a) and (c) initial and final permutation respectively

II. Passive Noninvasive Attack

The types of available power analysis are Simple power analysis (SPA), Differential power Analysis (DPA), Correlation power analysis (CPA), and higher-order analysis [9].

A. Simple Power Analysis (SPA)

SPA is a side-channel attack, which involves visual examination of graphs of the power used by a device over time. Variations in power consumption occur as the device performs different operations. In short, SPA exploits the relationship between the executed operations and the power leakage. For example, Fig 1 shows the power consumption of a DES algorithm. SPA shows sixteen identical pulses for rounds and two pulses for permutations and rotations.

B. Differential Power Analysis(DPA)

DPA is statistically analyzing power consumption measurements recorded from DUA. DPA inherently reduces the noise by its averaging technique [10]. The attacker does two steps. The first step, "recording", the power consumption of cryptosystem while plaintext is processed. The second step, "comparison", the recorded traces is compared with a power model of the DUA by correlation analysis. The result of correlation will detect the correct key; high peaks means correct key no peaks means false key guessing. DPA exploits the relationship between the processed data and the power leakage. The main advantage of DPA is no details required about the structure of the attacked algorithm while SPA shows only the type of the used algorithm. High-Order Differential Power Analysis (HO-DPA) is an advanced form of DPA attack. HO-DPA enables multiple data sources and different time offsets to be incorporated in the analysis.

C. Correlation Power Analysis(CPA)

In this technique, CPA is based on how a predicted power consumption model correlates with measured power consumption of DUA.

At first glance, the adversary builds power consumption model for DUA using Hamming Weight, Hamming Distance or any other models. These models target intermediate value dependent on key and plain message $f(P_i;K_S)$, where P_i is a known non-constant data value and K_S is a small part of the key. Consequently, plain messages P_i (i=1.....N) with every possible key K_S (S=1.....256) stimulate the selected power model and the results are recorded in (Nx256) predicted power matrix M_{pp}.

In the second part of attack, the same Plain messages P_i (i=1.....N) are encrypted by DUA meanwhile, the power consumption of DUA while the chip is operating is measured and recorded in in (NxT) measurement power matrix M_{mp}, where T denotes the length of the trace.

Finally, the analytical part, the correlation coefficient is the most common way to determine linear relationships between data. The correlation coefficient is used between M_{mp} and columns of M_{pp}. These results are recorded in a matrix of estimated correlation coefficients. An efficient way to compute this linear relation is to use Pearson coefficient that can be expressed as follows;

$$\rho M_{mp(t)}, M_{pp} = \frac{E\left(M_{mp}(t).M_{pp}\right) - E\left(M_{mp}(t)\right).E(M_{pp})}{\sqrt{var\left(M_{mp}(t)\right).var(M_{pp})}} \quad (1)$$

In this expression,

E(x) denotes the mean value of matrix x.
var(x) denotes the variance of matrix x.
$M_{mp}(t)$ denotes the measured power matrix at time "t".

The next Figure 2 expresses briefly CPA steps.

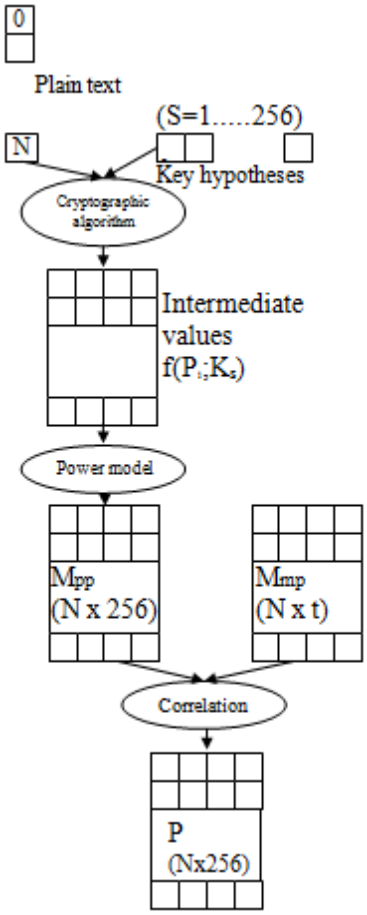


Fig. 2. Correlation Power Analysis

III. Types Of Noise Signals

Noise is an unwanted signal composed with genuine signal. According to cryptography, additive Noise to power signal have a great effect to immune side channel attack thus cryptography pursue to magnify the effect of noise on their implementation. There are two categories of noise: intentional

noise which is added by cryptography, normal noise which is any other type of noise.

When a measurement of a cryptographic device is repeated several times with constant input parameters, the resulting power traces are different. We refer to these fluctuations in the power traces as Electronic Noise [1]. Sources of this type of noise are varying from noise due to power supply, Quantization, and all other noise radiated from measurement setup components.

The device under attack consumed partial power relates to cipherkey which is important and rest of consumed power caused by cells that are not relevant for the attack as Switching Noise.

On other hand, Many cryptanalysts conduct researches and experiments to overcome the effect of the noise over attacking techniques, they could be summarized in three directions:

- Design of accurate power model to minimize Switching Noise.
- Use filters to minimize Electronic Noise.
- Conduct of Higher order statistics (HOS).

IV. Alignment Techniques

To prevent side channel attacks using power analysis techniques, cryptographers commonly implement DPA countermeasures that create misalignment in power trace sets and decrease the effectiveness of such attacks. On the other hand, Adversaries crucially work to realign the power traces.

There are two types of alignment: Static alignment and Dynamic alignment. Static misalignment is typically caused by inaccuracies in triggering the power measurements. Static alignment solves this problem by determining the duration of the timing inaccuracies, and shifting the traces accordingly [1].

In contrast, cryptographers actively use random time delays, varying clock frequencies and Random Process Interrupts (RPI). These techniques force cryptosystem sub-blocks to start operating at different and random times. Consequently, measured power consumptions are inherently misaligned. In these cases, static shifting cannot fully align the traces. Dynamic alignment is a general term for algorithms that match parts of several traces at different offsets, and perform nonlinear re-sampling of the traces.

A. Dynamic Time Warping

Dynamic Time Warping (DTW), was introduced to the data mining community by Berndt and Clifford [11]. In order to detect similar shapes with different phases, DTW method allows elastic shifting of the time axis. Also, speech processing community uses this technique for long time. The main drawback of DTW is processing time since performance on very large databases may be limited. Figure 3 demonstrates the power of DTW over traditional technique. DTW measures the distance between two sequences by elastically warping them in time.

Figure 4 shows the superiority of DTW over sliding window DPA (SW-DPA)[12]. SW-DPA is based on averaging fixed length clock cycles to restore the DPA peak in the face of random process interrupts.

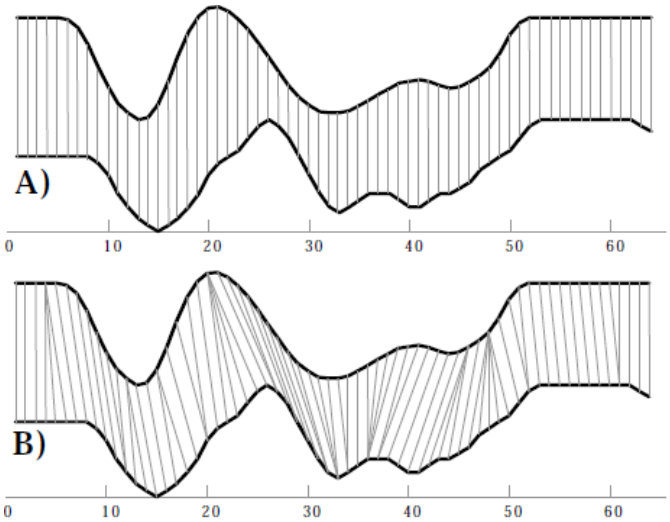


Fig. 3. Two sequences have an overall similar shape, they suffer from time misalignment, (A) i^{th} point on top sequence is aligned with the i^{th} point on the bottom that will produce a dissimilarity measure, (B) allows a more intuitive distance measure to be calculated.

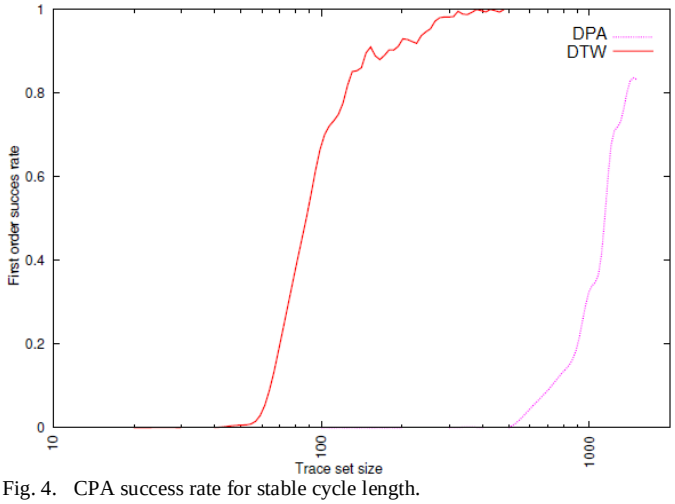


Fig. 4. CPA success rate for stable cycle length.

B. Correlation and Euclidean Alignment Technique

The main weakness of DTW is long processing time. In contrary to DTW, further techniques align power traces both partially and dynamically. Small portions of the traces (the interesting round part) are captured and aligned accordingly[13].

To save the processing time, this technique is based on extract the interesting round part from the whole power traces. This cutting concept limits later processing to shorter domain [14]. Figure 5 demonstrates the steps needed to align power traces.

Table I explains the steps of alignment. Fine tuning is considering with maximum similarity between selected round and other traces. This similarly is calculated by correlation, Euclidean [15].

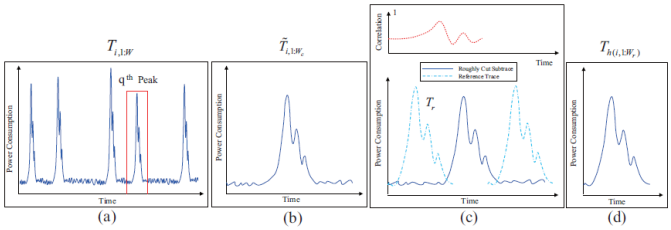


Fig. 5. Steps of aligning power traces (A)Find the qth peak, (B) Coarse Extraction, (C)Fine Tuning, and (D)Fine Tuned Extraction

TABLE I. STEPS OF ALIGNING POWER TRACES

Algorithm: Alignment Algorithm

Require: aligned power trace set

1. Find interesting round: Find the qth peak position (targeted round).
2. Coarse Extraction: Cut the selected round roughly to save all information belongs to that round.
3. Fine Tuning: Calculate maximum similarity by using Euclidean or Correlation to the selected round.
4. Fine Tuned Extraction: Cut the high similarity part in all traces.

TABLE II. MOMENTS OF DISTRIBUTION

Parameter	*Moment*	*Description*
Mean (μ)	1	Measure of central location.
Variance (σ2)	2	Measure of dispersion.
Skew	3	Measure of Asymmetry.
Kurtosis	4	Measure of peakedness.

The Euclidean distance or Euclidean metric is the distance between two vectors. The Euclidean distance is calculated by following equation:

$$d(RR,TR) = \sqrt{(RR_0 - TR_0)^2 + \cdots + (RR_n - TR_n)^2} \qquad (2)$$

Where:
RR: is the referenced vector.
TR: tested vector.

V. FOURTH ORDER CUMULANT

This research uses Higher Order Statistics (HOS) to find the moments of random signals that give the random variables some of their dominant features [16]. Moments are divided to two categories; Moments about the origin (raw moments) and Moments about the mean (central moment). Table II shows the first four distributions.

There are four ways to calculate the moments, they are:

1) Using the definition of moment.
2) Probability Generating Function [PGF].
3) Moment Generating Function [MGF].
4) Characteristic Function.

Moments are driven from the mathematical expectations of the random signal E{g(X)} [17].

The *raw moment*, is called the nth moment of X. It is denoted by n is given by

$$\mu'_n = E\{(X)^n\} = \sum_i (x_i^n) Px(x_i) \qquad (3)$$

Prime symbol is denoted to raw moment.

The central moments of random variable X are the moments of X with respect to its mean. Hence, the nth central moment of X, n, is defined as [17]

$$\mu_n = E\{(X-\mu)^n\} = \sum_i (x_i - \mu)^n Px(x_i) \qquad (4)$$

TABLE III. RAW AND CENTRAL MOMENTS

Moment	*Raw Moment*	*Central Moment*
μ_1	$E\{X\}=\mu$	0
$\mu 2$	$E\{X^2\}$	$E\{(X-\mu)^2\}=\sigma$
$\mu 3$	$E\{X^3\}$	$E\{(X-\mu)^3\}$
$\mu 4$	$E\{X^4\}$	$E\{(X-\mu)^4\}$

Table III contrasts between raw and central moments

Kurtosis measures the height and sharpness of the peak relative to the rest of the data. Higher values indicate a higher, sharper peak; lower values indicate a lower, less distinct peak. The kurtosis has no units: it's a pure number.

The normal distribution has a kurtosis of 3. Excess kurtosis is simply kurtosis-3[18].

- A normal distribution has kurtosis exactly 3 (excess kurtosis = 0). This is called mesokurtic.
- Any distribution has central peak is lower and broader, and its tails are shorter and thinner than normal distribution, its kurtosis < 3 (excess kurtosis < 0) is called platykurtic.
- Any distribution has central peak is higher and sharper, and its tails are longer and fatter than normal distribution, its kurtosis > 3 (excess kurtosis > 0) is called leptokurtic.

Kurtosis can be formally defined as a fourth population moment about the mean [19],

$$\beta_2 = \frac{E(X-\mu)^4}{(E(X-\mu)^2)^2} = \frac{\mu_4}{\sigma_4} \qquad (5)$$

Laplace was led to introduce a function known as a cumulative function, which is simply the logarithmic of the characteristic function [20]

VI. PROPOSED PREPROCESSING TECHNIQUE

The proposed idea aims to extract areas of interest using dynamic peak search [21]. Based on this technique, the 4th Cumulant method will duplicate its advantages to be used as

alignment technique. Furthermore, Cumulant shows superiority in:

- Eliminating the Gaussian noise.
- Treating misalignment.
- Reducing processing time.

We conducted two experiments to illustrate the power of the Cumulant method over the other traditional methods. The first experiment aims to compare Cumulant with other alignment techniques in absence of the noise. The second experiment highlights the advantages of Cumulant to prevent noise compared to traditional noise reduction methods with aligned power tracks.

A. *Cumulant vs. Alignment Techniques at Free Noise Environment*

In this part, free noise, misaligned measured power tracks are analyzed by the proposed idea and traditional alignment techniques. A comparison is made among these techniques to judge each one. Processing time, and differentiation between correct and false cipher keys will be the two parameters used to evaluate each technique.

Number of measured power tracks is frequently increased when alignment techniques are processed. Each time the difference between correlation results of correct and false cipherkeys is calculated and plotted. Figure 6 shows that, all traditional techniques including proposed one need the same number of power tracks to detect the correct cipherkey (approximately 81 power tracks). Over eighty one power tracks, the proposed idea and other techniques success to pickup the correct cipherkeys but in reality the traditional techniques show better results than the proposed idea.

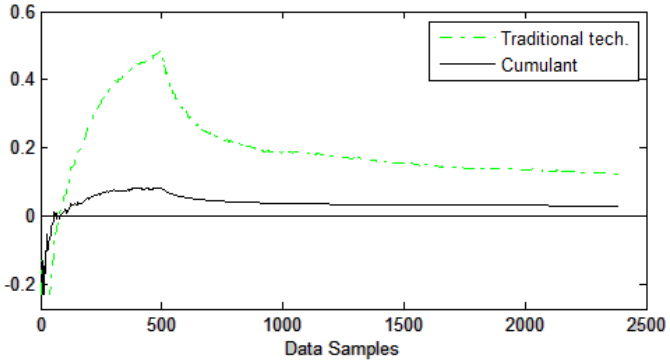


Fig. 6. Distinguishing between correct and false cipher keys.

Again, number of measured power tracks is frequently increased when alignment techniques are processed. Each time the processing time is calculated and plotted. Figure 7 shows the processing time plot for each technique.

Figure 7 illustrates that the processing time of traditional alignment techniques increases linearly with number of measured power tracks. Equations (6) and (7) express the rate of increasing of CPU processing time of traditional techniques. On other hand, It is obvious that the proposed idea keeps CPU processing time constant regardless to number of measured power tracks.

Moreover, CPU processing time of proposed idea (= $4x10^{-2}$sec) is 50% less than the lowest CPU processing time of traditional techniques.

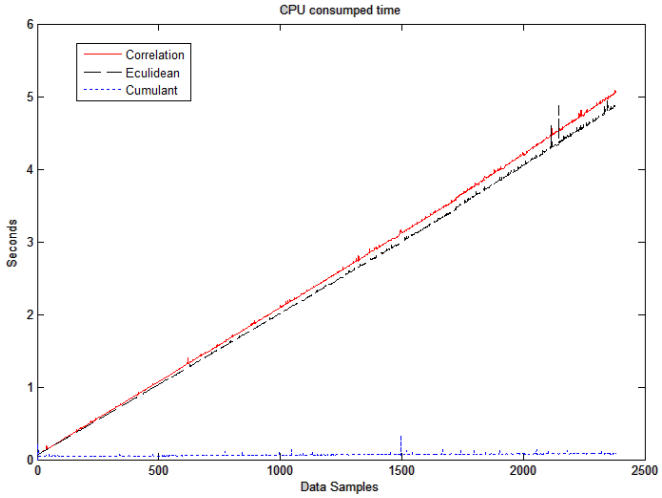


Fig. 7. The processing time of alignment techniques.

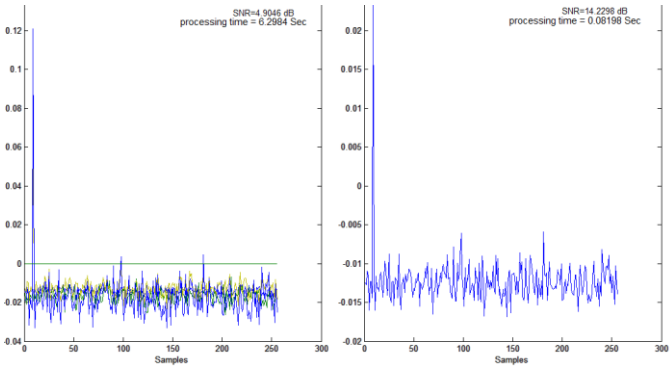


Fig. 8. A contrast between traditional alignment methods versus Cumulant.

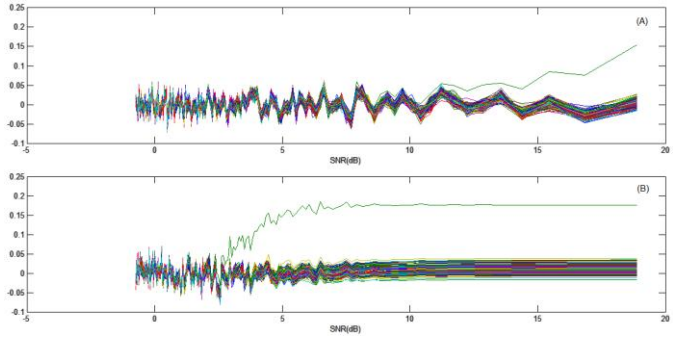


Fig. 9. RIJNDAEL SNR threshold (A)CPA behavior (B) Cumulant behavior.

$$\text{CPU time}|_{\text{Correlation}} = 2 \times 10^{-3}\text{tracks} + 7 \times 10^{-2} \quad (6)$$

$$\text{CPU time}|_{\text{Eculedian}} = 19 \times 10^{-4}\text{tracks} + 7 \times 10^{-2} \quad (7)$$

B. *Cumulant vs. Alignment Techniques at Noisy Environment*

Figure 8 shows a comparison between traditional alignment methods versus Cumulant. It's obvious that the correlation peak of true cipher key to false ones ("hint: we will denote the correlation peak of true cipher key to false ones as SNR_{corr}) of the proposed idea is enhanced three times the traditional

methods. Moreover, the speed of processing is 76 times higher than traditional methods.

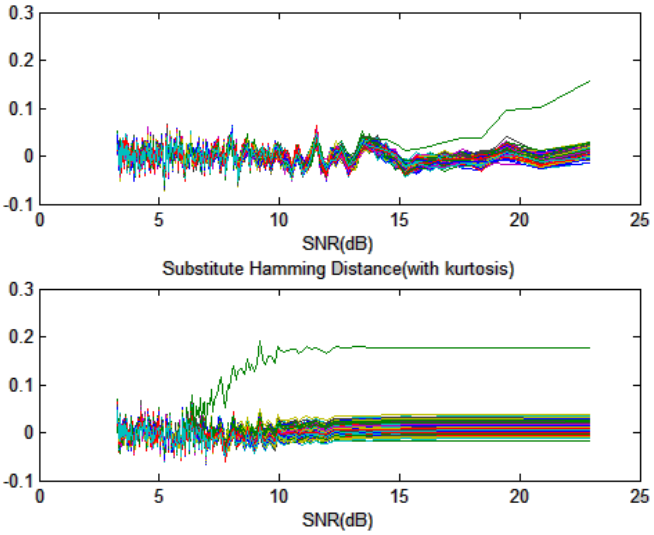


Fig. 10. Twofish SNR threshold (A)CPA behavior (B) Cumulant behavior.

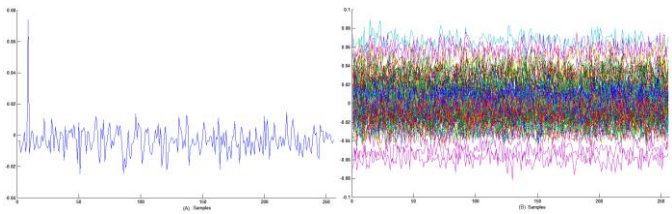


Fig. 11. Attacking RPI noisy stem by: (A)Cumulant (B) CPA.

Cumulant method demonstrates a great immunity to the noise, since it successfully attack a noisy system with SNR_{corr} = 2.5dB. This SNR_{corr} is reduced than those traditional methods by 75%.

Figure 9 shows the SNR limitation of Cumulant and traditional CPA. In Addition,Cumulant method is generic and independent of any specific cryptographic algorithm. The same results are obtained when applied to Twofish algorithm. Figure 10 shows the SNR limitation of Cumulant and traditional CPA when using Twofish algorithm.

It is obvious, that in case of a double impact system (noisy and RPI) all previous methods fails to get the cipherkey independently. As a result, combined techniques must be conducted and thus the processing time will be increased. Figure 11 shows the results of RPI noisy system exposed to attack by Cumulant and traditional methods. Figure 11 (A) demonstrates that Cumulant successfully attacks the system unlike the other.

VII. Conclusion

In this paper, we have proposed new features of fourth-order cumulant. These features include overcoming the noise added to the processed data, aligning measured tracks, speeding up the processing time to open new hopes to attack unbreakable algorithms due to time processing barriers. We have given the theoretical evaluation based on SNR criteria. The formulas to calculate these parameters have been given under a general form with flexible parameters, such as the noise level, and the number of side channel signals.

VIII. Future Work

Although the great results that are achieved by this paper, there are other noise types still affect the efficiency of SCA. These noise types need to be addressed and reduce its effects.

Extra studies are needed to highlight the benefits of the proposed algorithm over correlative noise countermeasures. Moreover, the limitation of the algorithm toward such noise would be calculated.

References

[1] S. Mangard, E. Oswald, and T. Popp, Power Analysis Attacks: Revealingthe Secrets of Smart Cards. Springer, 2007.

[2] S. P. Skorobogatov, Semi-invasive attacks a new approach to hardwaresecurity analysis. University of Cambridge Computer Laboratory,2005. [Online]. Available: http://www.cl.cam.ac.uk/TechReports/

[3] Implementation of Security in Actel Antifuse FPGAs. Alctel, ApplicationNote AC168, 2002.

[4] L. Hemme, A Differential Fault Analysis Against Early Rounds of Triple-DES. Proc. CHES, LNCS, vol. 3156, 2004.

[5] D. C. Y. K. JeaHoon Park, SangJae Moon and J. Ha, DifferentialFault Analysis for Round-Reduced AES by Fault Injection. ETRIJournal,Volume 33, Number 3, 2011.

[6] W. W. H. Chen and D. Feng, Differential Fault Analysis on CLEFIA.Proc. ICICS, LNCS, vol. 4861, 2007.

[7] D. G. W. Li and J. Li, Differential Fault Analysis on the ARIA Algorithm.Information Sciences, Elsevier, vol. 178, no. 19, 2008.

[8] J. J. P. Kocher and B. Jun, Differential Power Analysis. Proceedings of Crypto99, lecture notes in computer science, Vol. 1666, Springer-Verlag,1999.

[9] "updated time:june 2013, accessed date: Nov 2013," Wikipedia:the free encyclopedia.

[10] C. Chen, X. Li, L. Wu, and X. Zhang, "Design and implementation of a differential power analysis system for cryptographic devices," in Solid-State and Integrated Circuit Technology (ICSICT),, 2010.

[11] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 2, p. 143165, 1978.

[12] B. B. Jasper G. J. van Woudenberg, Marc F. Witteman, "Improving differential power analysis by elastic alignment," Cryptographers Track of the RSA Conference (CT-RSA), pp. 104–119, 2011.

[13] S. A. H. Qizhi Tian, "On clock frequency effiects in side channel attacks of symmetric block ciphers," IEEE Int. Conf. on New Technologies, Mobility and Security, 2012.

[14] M. S. Qizhi Tian, Abdulhadi Shoufan and S. A. Huss, "Power trace alignment for cryptosystems featuring random frequency countermeasures," Technical Report, Dept. of Computer Science, TU Darmstadt, 2012.

[15] S. A. H. Qizhi Tian, "On the attack of misaligned traces by power analysis methods," Technical Report, Dept. of Computer Science, TU Darmstadt, 2012.

[16] "updated time: 2012, accessed date: Nov 2013," Statistics Help.

[17] W. J. Stewart, "Probability, markov chains, queues, and simulation: The mathematical basis of performance modeling," in Princeton University Press. p. 105. ISBN 978-1-4008-3281-1, 2011.

[18] S. Brown, "Measures of Shape: Skewness and Kurtosis kernel description," http://www.tc3.edu/instruct/sbrown/stat/shape.htm, accessed: 2013-09-30.

[19] L. T. DeCarlo, "On the meaning and use of kertosis," Psychological Methods, vol. 2, pp. 292–307, 1997.

[20] R. A. F. E. A. Cornish, "Moments and cumulants in the specification of distriputions," revue de l'institute International de Statistique, vol. 5, pp. 307–320, 1937.

[21] S. A. H. Qizhi Tian, "A general approach to power trace alignment for the assessment of side-channel resistance of hardened cryptosystems," Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2012.

Quantum Cost Optimization for Reversible Sequential Circuit

Md. Selim Al Mamun
Jatiya Kabi Kazi Nazrul Islam University
Trishal, Mymensingh-2220, Bangladesh

David Menville
Pascack Valley High School
New Jersey, United States of America

***Abstract*—Reversible sequential circuits are going to be the significant memory blocks for the forthcoming computing devices for their ultra low power consumption. Therefore design of various types of latches has been considered a major objective for the researchers quite a long time. In this paper we proposed efficient design of reversible sequential circuits that are optimized in terms of quantum cost, delay and garbage outputs. For this we proposed a new 3*3 reversible gate called SAM gate and we then design efficient sequential circuits using SAM gate along with some of the basic reversible logic gates.**

Keywords—Flip-flop; Garbage Output; Reversible Logic; Quantum Cost

I. Introduction

In recent years, reversible computing has emerged as a promising technology. The primary reason for this is the increasing demands for lower power devices. In the early 1960s R. Landauer [1] demonstrated that losing bits of information causes loss of energy. Information is lost when an input cannot be recovered from its output. In 1973 C. H. Bennett [2] showed that energy dissipation problem can be avoided if the circuits are built using reversible logic gates.

Reversible logic has the feature to generate one to one correspondence between its input and output. As a result no information is lost and there is no loss of energy [3]. Although many researchers are working in this field, little work has been done in the area of sequential reversible logic. In the current literature on the design of reversible sequential circuits, the number of reversible gates is used as a major metric of optimization [4]. The number of reversible gates is not a good metric of optimization as reversible gates are of different type and have different quantum costs [5]. In this paper, we presented new designs of reversible sequential circuits that are efficient in terms of quantum cost, delay and the number of garbage outputs.

This paper is organized as follows: Section 2 presents some basic definitions related to reversible logic. Section 3 describes some basic reversible logic gates and their quantum implementation. Section 4 introduces our proposed gate 'Selim Al Mamun' (SAM) gate. Section 5 describes the logic synthesis of sequential circuits and comparisons with other researchers. Finally this paper is concluded with the Section 6.

II. Basic Definitions

In this section, some basic definitions related to reversible logic are presented. We formally define reversible gate, garbage output, delay in reversible circuit and quantum cost of reversible in reversible circuit.

A. Reversible Gate

A Reversible Gate is a k-input, k-output (denoted by k*k) circuit that produces a unique output pattern for each possible input pattern [6]. If the input vector is Iv where $Iv = (I_{1,j}, I_{2,j}, I_{3,j}, \ldots, I_{k-1,j}, I_{k,j})$ and the output vector is Ov where $Ov = (O_{1,j}, O_{2,j}, O_{3,j}, \ldots, O_{k-1,j}, O_{k,j})$, then according to the definition, for each particular vector j, $Iv \leftrightarrow Ov$.

B. Garbage Output

Every gate output that is not used as input to other gates or as a primary output is garbage. Unwanted or unused outputs which are needed to maintain reversibility of a reversible gate (or circuit) are known as Garbage Outputs. The garbage output of Feynman gate [7] is shown in Fig. 1 with *.

C. Delay

The delay of a logic circuit is the maximum number of gates in a path from any input line to any output line. The definition is based on two assumptions: (i) Each gate performs computation in one unit time and (ii) all inputs to the circuit are available before the computation begins.

In this paper, we used the logical depth as measure of the delay proposed by Mohammadi and Eshghi [8]. The delay of each 1x1 gate and 2x2 reversible gate is taken as unit delay 1. Any 3x3 reversible gate can be designed from 1x1 reversible gates and 2x2 reversible gates, such as CNOT gate, Controlled-V and Controlled-V^+ gates (V is a square-root-of NOT gate and V^+ is its hermitian). Thus, the delay of a 3x3 reversible gate can be computed by calculating its logical depth when it is designed from smaller 1x1 and 2x2 reversible gates.

D. Quantum Cost

The quantum cost of a reversible gate is the number of 1x1 and 2x2 reversible gates or quantum gates required in its design. The quantum costs of all reversible 1x1 and 2x2 gates are taken as unity [9]. Since every reversible gate is a combination of 1 x 1 or 2 x 2 quantum gate, therefore the quantum cost of a reversible gate can be calculated by counting the numbers of NOT, Controlled-V, Controlled-V^+ and CNOT gates used.

III. Quantum Analysis of Different Reversible Gates

Every reversible gate can be calculated in terms of quantum cost and hence the reversible circuits can be measured in terms

of quantum cost. Reducing the quantum cost from reversible circuit is always a challenging one and works are still going on in this area. This section describes some popular reversible gates and quantum equivalent diagram of each reversible gate.

A. Feynman Gate

Let I_v and O_v are input and output vector of a 2*2 Feynman gate where I_v and O_v are defined as follows: I_v = (*A, B*) and O_v = (*P = A, Q = A* ⊕ *B*). The quantum cost of Feynman gate is 1. The block diagram and equivalent quantum representation for a 2*2 Feynman gate are shown in Fig. 1.

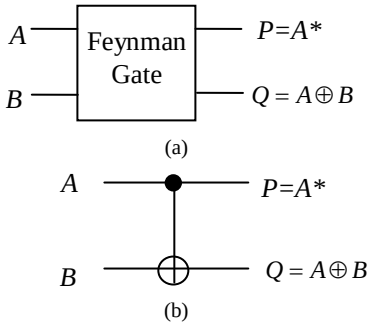


Fig. 1. (a) Block diagram of 2x2 Feynman gate and (b) Equivalent quantum representation

B. Double Feynman Gate

Let I_v and O_v are input and output vector of a 3*3 Double Feynman gate (DFG) where I_v and O_v are defined as follows: I_v = (*A, B, C*) and O_v = (*P = A, Q = A* ⊕ *B, R = A* ⊕ *C*). The quantum cost of Double Feynman gate is 2 [10]. The block diagram and equivalent quantum representation for 3*3 Double Feynman gate are shown in Fig. 2.

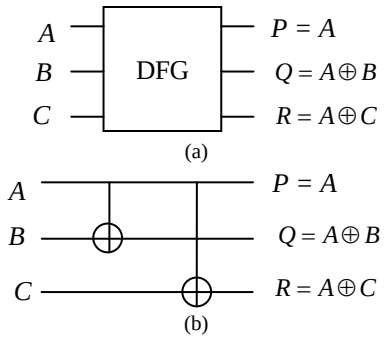


Fig. 2. (a) Block diagram of 3x3 Double Feynman gate and (b) Equivalent quantum representation.

C. Toffoli Gate

The input vector, I_v and output vector, O_v for 3*3 Toffoli gate (TG) [11] can be defined as follows: I_v = *(A, B, C)* and O_v = *(P = A, Q = B, R = AB*⊕ *C)*. The quantum cost of Toffoli gate is 5.

The block diagram and equivalent quantum representation for 3*3 Toffoli gate are shown in Fig. 3.

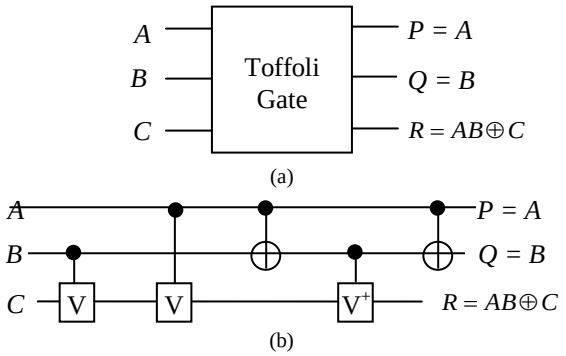


Fig. 3. (a) Block diagram of 3*3 Toffoli gate and (b) Equivalent quantum representation.

D. Frekdin Gate

The input vector, I_v and output vector, O_v for 3*3 Fredkin gate (FRG) [12] can be defined as follows: I_v = *(A, B, C)* and O_v = ($P = A$, $Q = \bar{A}B \oplus AC$, $R = \bar{A}C \oplus AB$). The quantum cost of Frekdin gate is 5. The block diagram and equivalent quantum representation for 3*3 Fredkin gate are shown in Fig. 4.

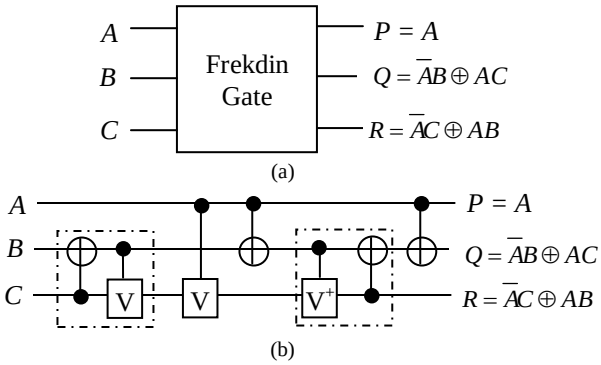


Fig. 4. (a) Block diagram of 3*3 Frekdin gate and (b) Equivalent quantum representation

E. Peres Gate

The input vector, I_v and output vector, O_v for 3*3 Peres gate (PG)[13] can be defined as follows: I_v = (*A, B, C*) and O_v = (*P = A, Q = A* ⊕ *B, R = AB* ⊕ *C*). The quantum cost of Peres gate is 4. The block diagram and equivalent quantum representation for 3*3 Peres gate are shown in Fig. 5.

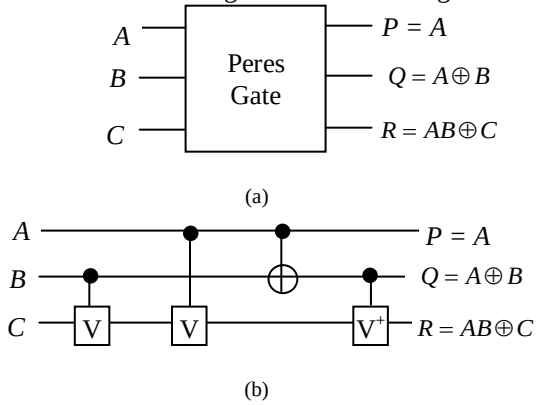


Fig. 5. Block diagram of 3*3 Peres and (b) Equivalent quantum representation

IV. PROPOSED SAM GATE

After The input vector, I_v and output vector, O_v for 3*3 SAM Gate is defined as follows: I_v = *(A, B, C)* and O_v = ($P = \bar{A}$, $Q = \bar{A}B \oplus A\bar{C}$, $R = \bar{A}C \oplus AB$). The block diagram of a 3*3 SAM gate is shown in Fig. 6**.**

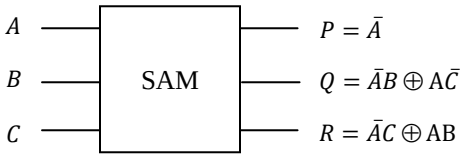


Fig. 6. Block diagram of a 3*3 SAM gate

The truth table for a 3x3 SAM gate is shown in Table I.

TABLE I. TRUTH TABLE FOR 3*3 SAM GATE

A	B	C	$P = \bar{A}$	$Q = \bar{A}B \oplus A\bar{C}$	$R = \bar{A}C \oplus AB$
0	0	0	1	0	0
0	0	1	1	0	1
0	1	0	1	1	0
0	1	1	1	1	1
1	0	0	0	1	0
1	0	1	0	0	0
1	1	0	0	1	1
1	1	1	0	0	1

We can verify from the corresponding truth table of the SAM gate that the output and input vectors have one to one mapping between them which satisfies the condition of reversibility of a gate. We can see from Table I that the 8 different input and output vectors unique means they have one to one mapping them. So the proposed gate satisfies the condition of reversibility.

The Equivalent quantum representation of the SAM gate and minimization of quantum cost are shown in Fig 7(a) through 7(d). The quantum cost of SAM gate is 4.

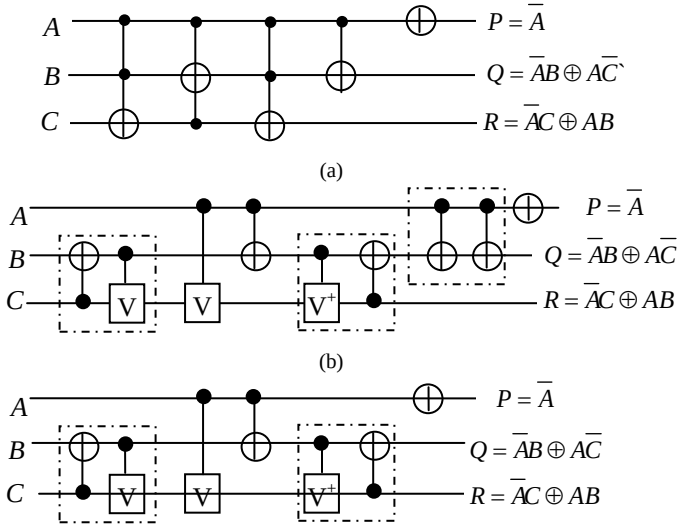


Fig. 7. Quantum cost of proposed SAM gate.

If we give 0 to 3rd input then we get NOT of 1st input in 1st output, OR or 1st and 2nd inputs in 2nd output and AND of 1st and 2nd inputs in 3rd output. This operation is shown in Fig. 8. So this gate can be used as two input universal gate.

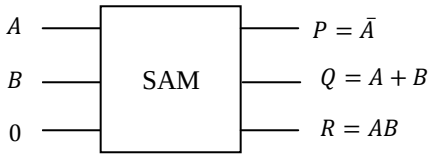


Fig. 8. SAM gate as two input universal gate.

V. DESIGN AND SYNTHESIS OF REVERSIBLE SEQUENTIAL CIRCUITS

In this section, we presented novel designs of reversible flip-flops that are optimized in terms of quantum cost, delay and garbage outputs.

A. The SR Flip-Flop

For SR flip-flop we modified the Peres gate. The modified Peres gate is shown in figure 9.

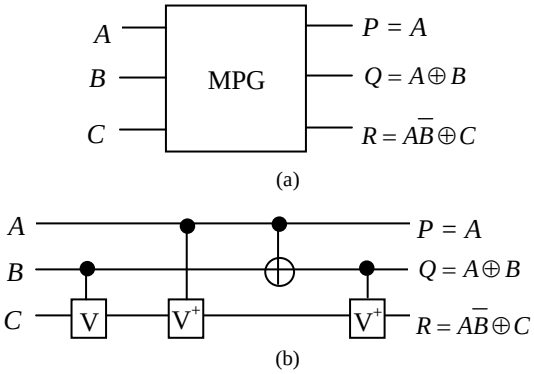


Fig. 9. (a) Block diagram of 3*3 MPG and (b) Equivalent quantum represenation

The characteristic equation of SR flip-flop is $Q = S + \overline{R}Q$. The SR flip-flop can be realized by a modified Peres gate (MPG). It can be mapped with the MPG by giving Q, R and S respectively in 1st, 2nd and 3rd inputs of the MPG. Fig. 10 shows the proposed design of SR flip-flop with Q and $\overline{Q}$ outputs.

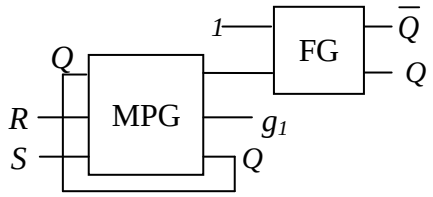


Fig. 10. Proposed design of SR flip-flop with Q and $\overline{Q}$ outputs

The proposed SR flip-flop with Q and $\overline{Q}$ outputs has quantum cost 5, delay 5 and has the bare minimum of 1 garbage bit. The proposed design of SR flip-flop achieves improvement ratios of 50% in terms of quantum cost, delay and garbage outputs compared to the design presented by Rice 2008 [14]. The improvement ratios compared to the design presented in Thapliyal et al.2010 [15] are 37%, 37% and 50% in terms quantum cost, delay and garbage outputs. The comparisons of our SR flip-flop (with Q and $\overline{Q}$ outputs) design with existing designs in literature are summarized in Table II.

TABLE II. COMPARISONS OF DIFFERENT TYPES OF SR FLIP-FLOPS WITH Q AND $\overline{Q}$ OUTPUTS

SR flip-flop design	Cost Comparison		
	Quantum Cost	*Delay*	*Garbage Outputs*
Proposed	5	5	1
Existing [14]	10	10	2
Existing [15]	8	8	2
Improvement(%) w.r.t. [14]	50	50	50
Improvement(%) w.r.t. [15]	37	37	50

This SR flip-flop design does not have enable signal (clock) and hence is not gated in nature. We proposed a design of gated SR flip-flop that can be realized by one MPG gate, one SAM and one FG gate. Another FG is needed to copy and produce the complement of Q. So we used a DFG instead of two FGs. The proposed gated SR flip-flop is shown in Fig. 11.

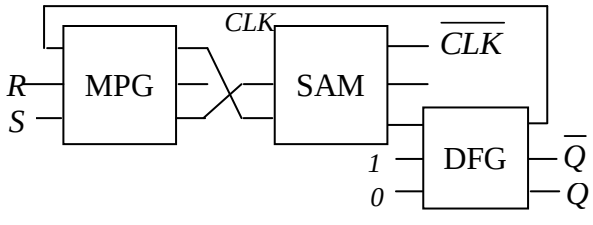


Fig. 11. Proposed design of gated SR flip-flop with Q and $\overline{Q}$ outputs

Proposed gated SR flip-flop with Q and $\overline{Q}$ outputs has quantum cost 10, delay 10 and has 2 garbage bits. The proposed design of gated SR flip-flop achieves improvement ratios of 41%, 41% and 33% in terms of quantum cost, delay and garbage outputs compared to the design presented in Thapliyal et al.2010 [15]. The comparisons of our gated SR flip-flop (with Q and $\overline{Q}$ outputs) design with existing designs in literature are summarized in Table III.

TABLE III. COMPARISONS OF DIFFERENT TYPES OF GATED SR FLIP-FLOPS WITH Q AND $\overline{Q}$ OUTPUTS

Gated SR flip-flop design	Cost Comparison		
	Quantum Cost	*Delay*	*Garbage Outputs*
Proposed	11	11	2
Existing[15]	17	17	3
Improvement in (%) w.r.t. [15]	41	41	33

Our proposed Master Slave SR flip-flop with only Q output is shown in Fig. 12.

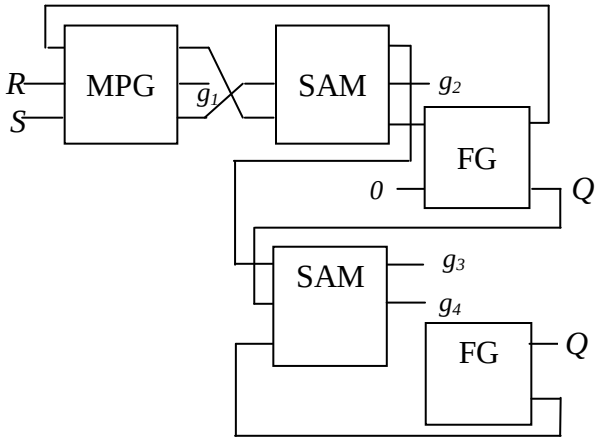


Fig. 12. Proposed Design of Master Slave SR flip-flop with only Q output

The proposed master-slave SR flip-flop with only Q output has quantum cost 14, delay 14 and has 4 garbage bits. The proposed design of master slave SR flip-flop achieves improvement ratios of 36% and 36% in terms of quantum cost and delay compared to the design presented in Thapliyal et al. 2010[15]. The comparisons of our Master Slave SR flip-flop design with existing designs in literature are summarized in Table IV.

TABLE IV. COMPARISONS OF DIFFERENT TYPES OF MASTER SLAVE SR FLIP-FLOPS WITH ONLY Q OUTPUT

Master slave SR flip-flop design	Cost Comparison		
	Quantum Cost	*Delay*	*Garbage Outputs*
Proposed	15	15	4
Existing[15]	22	22	4
Improvement in (%) w.r.t. [15]	36	36	0

B. The JK Flip-Flop

The characteristic equation of a JK flip-flop is $Q = J\overline{Q} + Q\overline{K}$. The JK flip-flop is realized by one SAM gate. It can be

mapped with the SAM gate by giving *Q*, *J* and *K* to 1st, 2nd and 3rd inputs to the SAM gate. The proposed JK flip-flop with Q and $\overline{Q}$ outputs is shown in Fig.13.

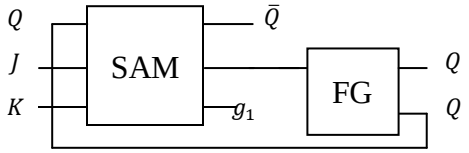


Fig. 13. Proposed design of JK flip-flop with Q and $\overline{Q}$ outputs

The proposed JK flip-flop with Q and $\overline{Q}$ outputs has quantum cost 5, delay 5 and has the bare minimum of 1 garbage bit. The proposed design of JK flip-flop achieves improvement ratios of 62%, 62% and 67% in terms of quantum cost, delay and garbage outputs compared to the design presented in Thapliyal et al. 2010[15]. The improvement ratios compared to the design presented in Lafifa Jamal et al. 2012[16] are 58%, 58% and 67% in terms quantum cost, delay and garbage outputs. The comparisons of our JK flip-flop (with Q and $\overline{Q}$ outputs) design with existing designs in literature are summarized in Table V.

TABLE V. COMPARISONS OF DIFFERENT TYPES OF JK FLIP-FLOPS WITH Q AND $\overline{Q}$ OUTPUTS

JK flip-flop design	Cost Comparisons		
	Quantum Cost	*Delay*	*Garbage Outputs*
Proposed	5	5	1
Existing[15]	13	13	3
Existing[16]	12	12	3
Improvement in (%) w.r.t. [15]	62	62	67
Improvement in (%) w.r.t. [16]	58	58	67

The characteristic equation of gated JK flip-flop is $Q = \overline{CLK}Q + CLK(J\overline{Q} + Q\overline{K})$. The gated JK flip-flop with Q and $\overline{Q}$ outputs is realized by two SAM gates and one DFG. The proposed gated JK flip-flop is shown in Fig.14.

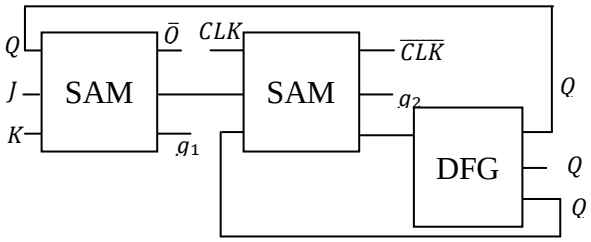


Fig. 14. Proposed Design of gated JK flip-flop with Q and $\overline{Q}$ outputs.

The proposed gated JK flip-flop with Q and $\overline{Q}$ outputs has quantum cost 10, delay 10 and has 2 garbage bits. The proposed design of gated JK flip-flop achieves improvement ratios of 37%, 37% and 33% in terms of quantum cost, delay and garbage outputs compared to the design presented in Thapliyal and Vinod 2007[17]. The improvement ratios compared to the design presented in Thapliyal et al. 2010[15] are 23%, 23% and 33% in terms quantum cost, delay and garbage outputs. The comparisons of our gated JK flip-flop (with Q and $\overline{Q}$ outputs) design with existing designs in literature are summarized in Table VI.

TABLE VI. COMPARISONS OF DIFFERENT TYPES OF GATED JK FLIP-FLOPS WITH Q AND $\overline{Q}$ OUTPUTS

Gated JK flip-flop design	Cost Comparisons		
	Quantum Cost	*Delay*	*Garbage Outputs*
Proposed	10	10	2
Existing[17]	16	16	3
Existing[15]	13	13	3
Improvement in (%) w.r.t. [17]	37	37	33
Improvement in (%) w.r.t. [15]	23	23	33

Our proposed Master Slave JK flip-flop with Q and $\overline{Q}$ outputs is shown in Fig.15.

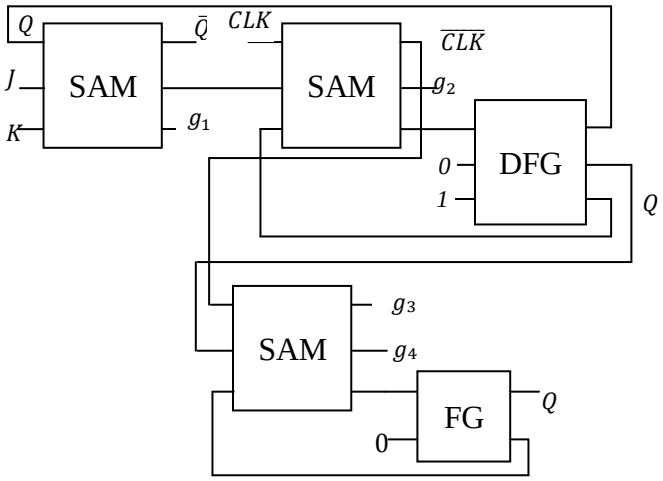


Fig. 15. Proposed Master Slave JK flip-flop with Q and $\overline{Q}$ outputs

The proposed master-slave JK flip-flop with Q and $\overline{Q}$ outputs has quantum cost 15, delay 15 and has 4 garbage bits. The proposed design of master slave JK flip-flop achieves improvement ratios of 37% and 37% in terms of quantum cost and delay compared to the design presented in Thapliyal and Vonod 2007[17]. The improvement ratios compared to the design presented in Thapliyal et al. 2010[15] are 21% and 21% in terms quantum cost and delay. The comparisons of our master slave JK flip-flop (with Q and $\overline{Q}$ outputs) design with existing designs in literature are summarized in Table VII.

TABLE VII. COMPARISONS OF DIFFERENT TYPES OF MASTER SLAVE JK FLIP-FLOPS WITH Q AND $\overline{Q}$ OUTPUTS

Master slave JK flip-flop design	Cost Comparisons		
	Quantum Cost	*Delay*	*Garbage Outputs*
Proposed	15	15	4
Existing[17]	24	23	5
Existing[15]	19	19	4
Improvement in (%) w.r.t. [17]	37	37	20
Improvement in (%) w.r.t. [15]	21	21	0

C. The D Flip-Flop

The characteristic equation of gated D flip-flop is $Q = \overline{CLK}.Q + CLK.D$. The D flip-flop can be realized by one SAM gate and one DFG. It can be mapped with SAM gate by giving CLK, D and Q respectively in 1st, 2nd and 3rd inputs of SAM gate. The Fig. 16 shows our proposed gated D flip-flop with with Q and $\overline{Q}$ outputs.

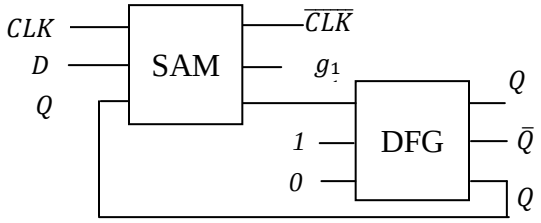


Fig. 16. Proposed design gated D flip-flop with Q and $\overline{Q}$ outputs.

The proposed gated D flip-flop with Q and $\overline{Q}$ outputs has quantum cost 6, delay 6 and has the bare minimum of 1 garbage bit. The proposed design of gated D flip-flop achieves improvement ratios of 14%, 14% and 50% in terms of quantum cost, delay and garbage outputs compared to the design presented in Thapliyal et al. 2010[15] and Lafifa Jamal et al. 2012[16]. The comparisons of our gated D flip-flop (with Q and $\overline{Q}$ outputs) design with existing designs in literature are summarized in Table VIII.

TABLE VIII. COMPARISONS OF DIFFERENT TYPES OF GATED D FLIP-FLOPS WITH Q AND $\overline{Q}$ OUTPUTS

D flip-flop design	Cost Comparisons		
	Quantum Cost	*Delay*	*Garbage Outputs*
Proposed	6	6	1
Existing[15]	7	7	2
Existing[16]	7	7	2
Improvement in (%) w.r.t. [15]	14	14	50
Improvement in (%) w.r.t. [16]	14	14	50

Our proposed Master Slave D flip-flop with Q and $\overline{Q}$ outputs is shown in Fig. 17.

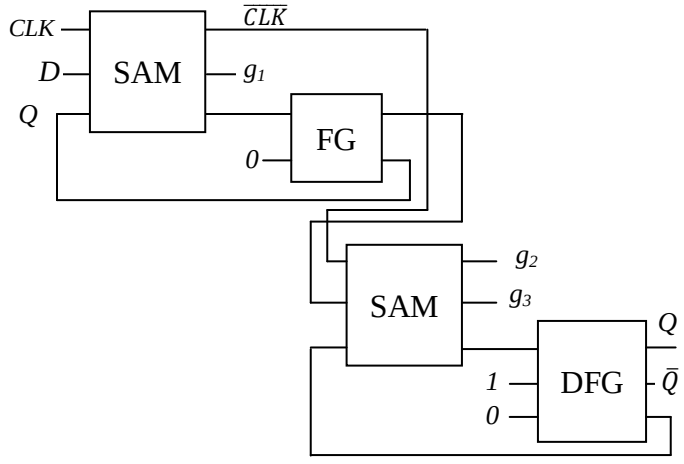


Fig. 17. Proposed design Master Slave D flip-flop with Q and $\overline{Q}$ outputs.

The proposed master-slave D flip-flop with Q and $\overline{Q}$ outputs has quantum cost 11, delay 11 and has 3 garbage bits. The proposed design of master slave D flip-flop achieves improvement ratios of 21% and 21% in terms of quantum cost and delay compared to the design presented in Chuang et al. 2008[18]. The improvement ratios compared to the design presented in Thapliyal et al. 2010[15] is 21% and 21% in terms quantum cost and delay. The comparisons of our master slave D flip-flop (with Q and $\overline{Q}$ outputs) design with existing designs in literature are summarized in Table IX.

TABLE IX. COMPARISONS OF DIFFERENT TYPES OF MASTER SLAVE D FLIP-FLOPS WITH Q AND $\overline{Q}$ OUTPUTS

Master Slave D flip-flop design	Cost Comparisons		
	Quantum Cost	*Delay*	*Garbage Outputs*
Proposed	11	11	3
Existing[18]	14	14	3
Existing[15]	13	13	3
Improvement in (%) w.r.t. [17]	21	21	0
Improvement in (%) w.r.t. [15]	15	15	0

VI. CONCLUSION

Reversible latches are going to be the main memory block for the forthcoming quantum devices. In this paper we proposed optimized reversible D latch and JK latches with the help of proposed SAM gates. Appropriate algorithms and theorems are presented to clarify the proposed design and to establish its efficiency. We compared our design with existing ones in literature which claims our success in terms of number of gates, number of garbage outputs and delay. This optimization can contribute significantly in reversible logic community.

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for their supports and constructive feedback, which helped significantly to improve technical quality of this paper.

REFERENCES

[1] Rolf Landauer, "Irreversibility and Heat Generation in the Computing Process", IBM Journal of Research and Development, vol. 5, pp. 183-191, 1961.

[2] Charles H.Bennett, "Logical Reversibility of computation", IBM Journal of Research and Development, vol. 17, no. 6, pp. 525-532, 1973.

[3] Perkowski, M., A.Al-Rabadi, P. Kerntopf, A. Buller, M. Chrzanowska-Jeske, A. Mishchenko, M. Azad Khan, A. Coppola, S. Yanushkevich, V. Shmerko and L. Jozwiak, "A general decomposition for reversible logic", Proc. RM'2001, Starkville, pp: 119-138, 2001

[4] J.E Rice, "A New Look at Reversible Memory Elements", Proceedings International Symposium on Circuits and Systems(ISCAS) 2006, Kos, Greece, May 21-24 ,2006, pp. 243-246.

[5] Dmitri Maslov and D. Michael Miller, "Comparison of the cost metrics for reversible and quantum logic synthesis", http://arxiv.org/abs/quant-ph/0511008, 2006

[6] Md. Selim Al Mamun and Syed Monowar Hossain. "Design of Reversible Random Access Memory." International Journal of Computer Applications 56.15 (2012): 18-23.

[7] Richard P.Feynman, "Quantum mechanical computers," Foundations of Physics, vol. 16, no. 6, pp. 507-531, 1986.

[8] Mohammadi,M. and Mshghi,M, On figures ofmerit in reversible and quantumlogic designs, Quantum Inform. Process. 8, 4, 297–318, 2009.

[9] D. Michael Miller, Dmitri Maslov, GerhardW. Dueck, A Transformation Based Algorithm for Reversible Logic Synthesis, Annual ACM IEEE Design Automation Conference,Proceedings of the 40th annual Design Automation Conference, Anaheim, CA, USA Pages: 318 – 323.

[10] Perkowski, M., "A hierarchical approach to computer-aided design of quantum circuits", 6th International Symposium on Representations and Methodology of Future Computing Technology, 201-209, 2003.

[11] Tommaso Toffoli, "Reversible Computing," Automata, Languages and Programming, 7th Colloquium of Lecture Notes in Computer Science, vol. 85, pp. 632-644, 1980.

[12] Edward Fredkin and Tommaso Toffoli, "Conservative Logic," International Journal of Theoretical Physics, vol. 21, pp. 219-253, 1982.

[13] A. Peres, "Reversible Logic and Quantum Computers," Physical Review A, vol. 32, pp. 3266-3276, 1985.

[14] J. E. Rice, An introduction to reversible latches. The Computer journal,Vol. 51, No.6, 700–709. 2008.

[15] Himanshu Thapliyal and Nagarajan Ranganathan, Design of Reversible Sequential Circuits Optimizing Quantum Cost, Delay, and Garbage Outputs, ACMJournal onEmerging Technologies inComputer Systems,Vol. 6,No. 4,Article 14, Pub. date:December 2010.

[16] Lafifa Jamal, Farah Sharmin, Md. Abdul Mottalib and Hafiz Md. Hasan Babu, Design and Minimization of Reversible Circuits for a Data Acquisition and Storage System, International Journal of Engineering and Technology Volume 2 No. 1, January, 2012

[17] H. Thapliyal and A. P. Vinod, "Design of reversible sequential elements with feasibility of transistor implementation" In Proc. the 2007 IEEE Intl. Symp. On Cir.and Sys., pages 625–628, New Orleans, USA, May 2007.

[18] M.-L. Chuang and C.-Y. Wang, "Synthesis of reversible sequential elements," ACM journal of Engineering Technologies in Computing Systems (JETC). Vol. 3, No.4, 1–19, 2008.

Color, texture and shape descriptor fusion with Bayesian network classifier for automatic image annotation

Mustapha OUJAOURA[1], Brahim MINAOUI[2], Mohammed FAKIR[3]
Laboratory of Information Processing and Telecommunications, Computer Science Department,
Faculty of Science and Technology, Sultan Moulay Slimane University
Mghila, PO Box. 523, Béni Mellal, Morocco

***Abstract*—Due to the large amounts of multimedia data prevalent on the Web, Some images presents textural motifs while others may be recognized with colors or shapes of their content. The use of descriptors based on one's features extraction method, such as color or texture or shape, for automatic image annotation are not efficient in some situations or in absence of the chosen type. The proposed approach is to use a fusion of some efficient color, texture and shape descriptors with Bayesian networks classifier to allow automatic annotation of different image types. This document provides an automatic image annotation that merges some descriptors in a parallel manner to have a vector that represents the various types of image characteristics. This allows increasing the rate and accuracy of the annotation system. The Texture, color histograms, and Legendre moments, are used and merged respectively together in parallel as color, texture and shape features extraction methods, with Bayesian network classifier, to annotate the image content with the appropriate keywords. The accuracy of the proposed approach is supported by the good experimental results obtained from ETH-80 databases.**

Keywords*—image annotation; k-means segmentation; Bayesian networks; color histograms; Legendre moments; Texture; ETH-80 database

I. INTRODUCTION

With the rapid development of Internet communication technology and digital imaging technology, users can easily get many networked digital information archives by a variety of ways. Searching this digital information archives on the Internet and elsewhere has become a significant part of our daily lives. Amongst the rapidly growing body of information, many digital images are not reached. The task of automated image retrieval is complicated by the fact that many images do not have suitable textual descriptions and annotations. Retrieval of images through analysis of their visual content is therefore an exciting and notable research challenge.

With regard to the long standing problem of the semantic gap between low-level image features and high-level human knowledge, the image retrieval community has recently shifted its emphasis from low-level features analysis to high-level image semantics extraction. Therefore, image semantics extraction is of great importance to content-based image retrieval because it allows the users to freely express what images they want. Semantic content annotation is the basis for semantic content retrieval. The automatically obtained keywords from image annotation process can be used to represent the images content to facilitate their retrieval.

Automatic object recognition and annotation are essential tasks in these image retrieval systems. Indeed, Annotated images play important role in information processing; they are useful for image retrieval based on keywords and image content management [1]. For that reason, many research efforts have aimed at annotating objects contained in visual streams. Image content annotation facilitates conceptual image indexing and categorization to assist text-based image search that can be semantically more significant than search in the absence of any text [2], [3].

Manual annotation is not only boring but also not practical in many cases, due to the abundance of information. Many images are therefore available without suitable textual annotation. Automatic image content annotation becomes a recent research interest [3], [4]. It attempts to explore the visual characteristics of images and associate them with image contents and semantics to use textual request for image retrieval and searching; automatic image annotation is an efficient technology for improving the image retrieval.

The rest of the paper is organized as follows. Firstly, the section 2 presents the proposed annotation system. The Section 3 discusses the image segmentation while the section 4 presents a brief formulation of color histograms, Texture, and Legendre moments as features extraction method. The Section 5 is reserved for the annotation by the approach of image classification using a fusion of several descriptors that are the color histograms, Texture, and Legendre moments with Bayesian network classifier. The Section 6 presents the experimental results of the image annotation based on the proposed approach. Finally, in the last section, the main conclusion concerning the proposed approach is given in addition to the possible future works.

II. ANNOTATION SYSTEM

Automatic image annotation consists of associating, to each image, a group of words that describes the visual contents of the image without human intervention. This task has been, and still, the subject of many studies [5], [6], [7], [8], [9], [10]. Several ways are used to deal with the problem of automatic image annotation. A recent review on automatic image

annotation techniques is presented in [11]. Using machine learning methods from examples of annotated images, many automatic image annotation techniques aim to learn the relationship between keywords and visual features. The learned relationships are then used to assign keywords to non-annotated images.

As an improvement of the previous works [12, 13, 14], this problem is considered for image that are described globally or not yet labeled with a suitable text terms. Some images on the web presents textural motifs, others can be recognized with colors or shapes of their content. The fusion of multiple descriptors, which are of different types such as color or texture or shape descriptors, can increase the effectiveness of images representation allowing their annotations in a manner that is more accurate than when using one descriptor type's. Indeed, in situations where images have a different descriptor type's from that used, the results will be catastrophic. Also, individual classifier and features descriptor results are limited or not suited for some situations [15]. So, their fusion is important, it can improve the annotation results and improve the accuracy of the annotation system. In such case, the proposed approach is to use a fusion of some efficient color, texture and shape descriptors with Bayesian network classifier to allow automatic annotation of different image types. The objective of this document is to provide an automatic image annotation that merges some descriptors in a parallel manner to have a features vector that represents the various types of image characteristics. This approach can allow increasing the rate and accuracy of the annotation system. The block diagram of the image annotation system adopted in this work is shown in Fig.1.

The system contains several phases. Firstly, the query image is segmented into regions that represent objects in the image, secondly, the features vector of each region are computed and extracted from the image, and those features are merged and are finally fed into input of the already trained classifiers that is the Bayesian Network to decide and choose the appropriate keywords for annotation tasks.

III. K-MEANS IMAGE SEGMENTATION

Usually, the features vector extracted from the entire image loses local information. Therefore, it is necessary to segment an image into regions or objects of interest and use of local characteristics. Image segmentation is a method that localizes and extracts an object from an image or divides the image into several regions. It plays important role in many applications for image processing, and still remains a challenge for scientists and researchers.

The efforts and attempts are still being made to improve the segmentation techniques. With the improvement of computer processing capabilities, several possible segmentation techniques of an image have emerged: threshold, region growing, k-means, active contours, level sets, etc...[16].

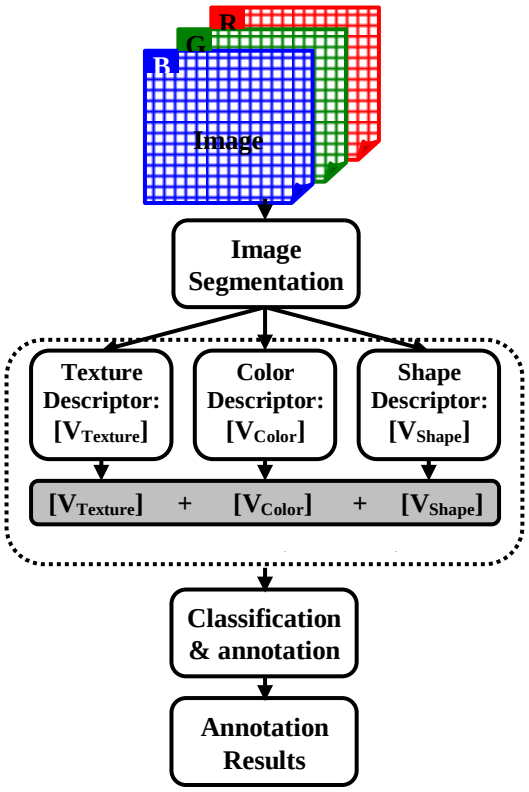


Fig. 1. Block Diagram of the proposed annotation system.

Among the segmentation methods, the k-means is well suited because of its simplicity and has been successfully used several times as a segmentation technique of digital images.

The K-Means algorithm is based on a clustering algorithm that does not require the presence of a learning database. So, this algorithm can organize the pixels of the image.

Given a set of image pixels $X = \{p_1, p_2, \ldots, p_n\} \in R^d$ where each pixel is a veritable vector of dimension d = 3 in the case of a colour image (d = 5 if the pixels coordinates are introduced as information of spatial coherence or connectivity). The k-Means algorithm aims to classify and divide the n pixels of the image into k sets or regions ($k \leq n$) $S = \{R_1, R_2, \ldots, R_k\}$ with a manner that minimizes the inter-class variance, that results in minimizing the sum of squared Euclidean distances among the clusters defined by:

$$E = \sum_{i=1}^{k} \sum_{p_j \in R_i} \left\| p_j - m_i \right\|^2 = \sum_{i=1}^{k} Card(R_i) \times Var(R_i) \quad (1)$$

Where:

- p_j pixel vector;

- $Card(R_i)$ is the number of pixels in the cluster or region R_i ;
- $m_i = \frac{\sum_{p_j \in R_i} p_j}{Card(R_i)}$ is the centre of the cluster or region R_i also known as kernel;
- $Var(R_i) = \frac{\sum_{p_j \in R_i} \|p_j - m_i\|^2}{(Card(R_i))^2}$ is the variance of pixel cluster or region.

The k-means image segmentation algorithm finds the pixels groups that minimize the quantity E defined above. This comes somehow for each cluster or region, to minimize the following quantity:

$$\sum_{p_j \in R_i} \|p_j - m_i\|^2 \quad (2)$$

The principle of the minimization algorithm of this error can result in the following main steps [18]:

1) Choosing the number of clusters (number of kernels);
2) Initialization of clusters and their kernels;
3) Updating clusters by optimizing the error clustering;
4) Calculation and revaluation of the new clusters;
5) Iterate and repeat steps 3 and 4 until clusters stabilization.

The number of clusters k can match approximately the number of dominant colors used to represent the image. The determination of k is done using the color histograms.

After transformation of the color image into a single image formed by the reduced numbers of colors, the cluster number k is selected to be the number of peaks in the histogram from the transformed image.

An example of image segmentation, by using K-means segmentation algorithm, is presented in Fig. 2.

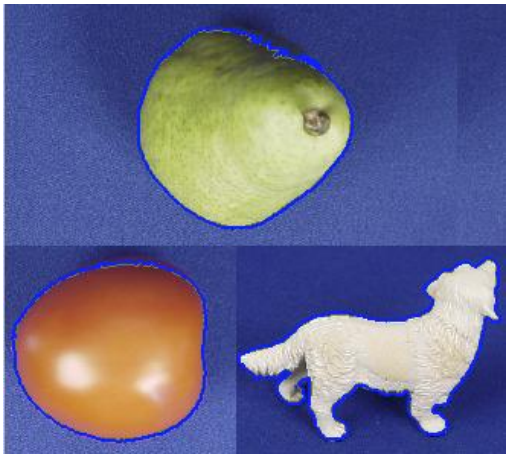


Fig. 2. Example of K-means image segmentation.

IV. Features Extraction

After dividing the original image into several distinct regions that correspond to objects in a scene, the feature vector must be extracted carefully from a region to reduce the rich content and large input data of images and preserve the content representation of the entire image. Therefore, the feature extraction task can decrease the processing time. it enhances not only the retrieval and annotation accuracy, but also the annotation speed as well, since a large image database can be organized according to the classification rule and, therefore, search can be done [19].

In the feature extraction method, the representation of the image content must be considered in some situations such as: translation, rotation and change of scale. This is the reason that justifies the use of color histograms and moments for feature extraction method from the segmented image.

All these features are extracted for all the images in reference database and stored with keywords in features database. For more precision and accuracy in the annotation system, they can be combined together and feed to the input of the classifier [15]. This combination costs more time for training the classifiers due to the size of the resulted features.

A. Color histogram

Typically, the color of an image is represented through some color model. There exist various color models to describe color information. The more commonly used color models are RGB (red, green, blue), HSV (hue, saturation, value) and Y, Cb, Cr (luminance and chrominance). Thus, the color content is characterized by 3 channels from some color models. In this paper, we used RGB color models. One representation of color image content is by using color histogram. Statistically, it denotes the joint probability of the intensities of the three color channels [20].

Color histogram describes the distribution of colors within a whole or within an interest region of image. The histogram is invariant to rotation, translation and scaling of an object but the histogram does not contain semantic information, and two images with similar color histograms can possess different contents.

The histograms are normally divided into bins to coarsely represent the content and reduce dimensionality of subsequent classification and matching phase. A color histogram H for a given image is defined as a vector by:

$$H = \left\{ \begin{array}{l} h[i \in \{1,\ldots,k\}] = \dfrac{\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} \delta(f(x,y) - C(i))}{M \times N} \\ / \ (i-1) \times E\left(\dfrac{256}{k}\right) \le C(i) < i \times E\left(\dfrac{256}{k}\right) \end{array} \right\} \quad (3)$$

Where:

- i represent a color in the color histogram;
- E(x) denotes the integer part of x;

- h[i] is the number of pixel with color i in that image;
- k is the number of bins in the adopted color model;
- And δ is the unit pulse defined by:

$$\delta(x, y) = \begin{cases} 1 & if \quad x = y = 0 \\ 0 & else \end{cases} \tag{4}$$

In order to be invariant to scaling change of objects in images of different sizes, color histograms H should be divided by the total number of pixels M x N of an image to have the normalized color histograms.

For a three-channel image, a feature vector is then formed by concatenating the three channel histograms into one vector.

B. Legendre Moments

In this paper, the Legendre moments are calculated for each one of the 3 channel in a color image. A feature vector is then formed by concatenating the three channel moments into one vector.

The Legendre moments [21] for a discrete image of M x N pixels with intensity function f(x, y) is the following:

$$L_{pq} = \lambda_{pq} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} P_p(x_i)\, P_q(y_j)\, f(x, y) \tag{5}$$

Where $\lambda_{pq} = \frac{(2p+1)(2q+1)}{M \times N}$, x_i and y_j denote the normalized pixel coordinates in the range of [-1, +1], which are given by:

$$\begin{cases} x_i = \dfrac{2x - (M-1)}{M-1} \\ y_j = \dfrac{2y - (N-1)}{N-1} \end{cases} \tag{6}$$

$P_p(x)$ is the p[th]-order Legendre polynomial defined by:

$$P_p(x) = \sum_{k=0}^{p} \left\{ \frac{(-1)^{\frac{p-k}{2}} (p+k)!\, x^k}{2^p\, k! \left(\frac{p-k}{2}\right)! \left(\frac{p+k}{2}\right)!} \right\}_{p-k=even} \tag{7}$$

In order to increase the computation speed for calculating Legendre polynomials, we used the recurrent formula of the Legendre polynomials defined by:

$$\begin{cases} P_p(x) = \dfrac{(2p-1)x}{p} P_{p-1}(x) - \dfrac{(p-1)}{p} P_{p-2}(x) \\ P_1(x) = x \quad , \quad P_0(x) = 1 \end{cases} \tag{8}$$

C. Texture Descriptors

Several images have textured patterns. Therefore, the texture descriptor is used as feature extraction method from the segmented image.

The texture descriptor is extracted using the co-occurrence matrix introduced by Haralick in 1973 [22]. So for a color image I of size $N \times N \times 3$ in a color space (C_1, C_2, C_3), for $(k, l) \in [1, \cdots, N]^2$ and $(a, b) \in [1, \cdots, G]^2$, the co-occurrence matrix $M_{k,l}^{C,C'}[I]$ of the two color components $C, C' \in \{C_1, C_2, C_3\}$ from the image I is defined by:

$$M_{k,l}^{C,C'}([I], a, b) = \frac{1}{(N-k)(N-l)} \times \sum_{i=1}^{N-k} \sum_{j=1}^{N-l} \delta(I(i, j, C) - a\,,\, I(i+k, j+l, C') - b) \tag{9}$$

Where δ is the unit pulse defined by:

$$\delta(x, y) = \begin{cases} 1 & if \quad x = y = 0 \\ 0 & else \end{cases} \tag{10}$$

Each image I in a color space (C_1, C_2, C_3) can be characterized by six color co-occurrence matrix.

Matrix $M^{C_2, C_1}[I]$, $M^{C_3, C_1}[I]$ and $M^{C_3, C_2}[I]$ are not taken into account because they can be deduced respectively by diagonal symmetry from matrix $M^{C_1, C_2}[I]$, $M^{C_1, C_3}[I]$ and $M^{C_2, C_3}[I]$. As they measure local interactions between pixels, they are sensitive to significant differences in spatial resolution between the images. To reduce this sensitivity, it is necessary to normalize these matrices by the total number of the considered co-occurrences matrix:

$$M_{k,l}^{C,C'}([I], a, b) = \frac{M_{k,l}^{C,C'}([I], a, b)}{\sum_{i=0}^{T-1} \sum_{j=0}^{T-1} M_{k,l}^{C,C'}([I], i, j)} \tag{11}$$

Where T is the number of quantization levels of the color components.

To reduce the large amount of information of these matrices, the 14 Haralick indices [22] of these matrices are used. There will be then 84 textures attributes for six co-occurrence matrices (14×6).

V. IMAGE CLASSIFICATION AND ANNOTATION

The goal of pattern classification is to allocate an object represented by a number of feature vectors into one of a finite set of classes from the reference database. In order to classify unknown patterns, a certain number of training samples available for each class are used to train the classifier. The learning task is to compute a classifier or model that approximates the mapping between the input-output examples and correctly labels the training set with some level of accuracy. This can be called the training or model generation stage. After the model is generated and trained, it is able to classify an unknown instance, into one of the learned class labels in the training set. More specifically, the classifier calculates the similarity of all trained classes and assigns the unlabeled instance to the class with the highest similarity measure.

Therefore, image annotation can be approached by the model or the classifier generated and trained to bridge the gap between low-level feature vectors and high-level concepts; a function is learned which can directly correspond the low-level feature sets to high-level conceptual classes. There are several types of classifier that can be used for classification. The Bayesian network classifier is used in this paper.

Bayesian networks are based on a probabilistic approach governed by Bayes' rule. The Bayesian approach is then based on the conditional probability that estimates the probability of occurrence of an event assuming that another event is verified. A Bayesian network is a graphical probabilistic model representing the random variable as a directed acyclic graph. It is defined by [22]:

- $G=(X, E)$, Where X is the set of nodes and E is the set of edges, G is a Directed Acyclic Graph (DAG) whose vertices are associated with a set of random variables $X=\{X_1, X_2, \cdots, X_n\}$;
- $\theta=\{P(X_i|Pa(X_i))\}$ is a conditional probabilities of each node X_i relative to the state of his parents $Pa(X_i)$ in G.

The graphical part of the Bayesian network indicates the dependencies between variables and gives a visual representation tool of knowledge more easily understandable by users. Bayesian networks combine qualitative part that are graphs and a quantitative part representing the conditional probabilities associated with each node of the graph with respect to parents [23].

Pearl and all [24] have also shown that Bayesian networks allow to compactly representing the joint probability distribution over all the variables:

$$P(X)=P(X_1, X_2, \cdots, X_n)=\prod_{i=1}^{n} P(X_i|Pa(X_i)) \quad (12)$$

Where $Pa(X_i)$ is the set of parents of node X_i in the graph G of the Bayesian network.

This joint probability could be actually simplified by the Bayes rule as follows [25]:

$$\begin{aligned} P(X) &= P(X_1, X_2, \cdots, X_n)=\prod_{i=1}^{n} P(X_i|Pa(X_i)) \\ &= P(X_n|X_{n-1}, \cdots, X_1)\times P(X_{n-1}|X_{n-2}, \cdots, X_1)\times\cdots\times P(X_2|X_1)\times P(X_1) \\ &= P(X_1)\times\prod_{i=2}^{n} P(X_i|X_{i-1}, \cdots, X_1) \end{aligned} \quad (13)$$

The construction of a Bayesian network consists in finding a structure or a graph and estimates its parameters by machine learning. In the case of the classification, the Bayesian network can have a class node C_i and many attribute nodes X_j. The naive Bayes classifier is used in this paper due to its robustness and simplicity. The Fig. 3 illustrates its graphical structure.

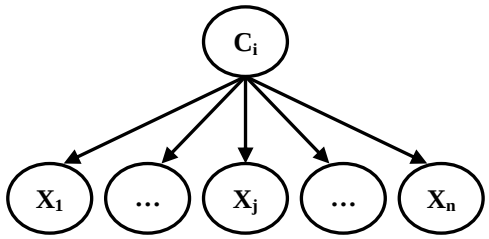


Fig. 3. Naive Bayes classifier structure.

To estimate the Bayesian network parameters and probabilities, Gaussian distributions are generally used. The conditional distribution of a node relative to its parent is a Gaussian distribution whose mean is a linear combination of the parent's value and whose variance is independent of the parent's value [26]:

$$P(X_i|Pa(X_i))=\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left\{\frac{-1}{2\sigma_i^2}\left(x_i-\left(\mu_i+\sum_{j=1}^{n_i}\frac{\sigma_{ij}}{\sigma_j^2}(x_j-\mu_j)\right)\right)^2\right\} \quad (14)$$

Where,

- $Pa(X_i)$ Are the parents of X_i;
- μ_i, μ_j, σ_i *and* σ_j are the means and variances of the attributes X_i and X_j respectively without considering their parents;
- n_i is the number of parents;

- σ_{ij} is the regression matrix of weights.

After the parameter and structure learning of a Bayesian network, The Bayesian inference is used to calculate the probability of any variable in a probabilistic model from the observation of one or more other variables. So, the chosen class C_i is the one that maximizes these probabilities [27], [28]:

$$P(C_i|X) = \begin{cases} P(C_i)\prod_{j=1}^{n} P(X_j|Pa(X_j), C_i) & \text{if } X_j \text{ has parents} \\ P(C_i)\prod_{j=1}^{n} P(X_j|C_i) & \text{else} \end{cases} \quad (15)$$

For the naive Bayes classifier, the absence of parents and the variables independence assumption are used to write the posterior probability of each class as given in the following equation [29]:

$$P(C_i|X) = P(C_i)\prod_{j=1}^{n} P(X_j|C_i) \quad (16)$$

Therefore, the decision rule d of an attribute X is given by:

$$\begin{aligned} d(X) &= \arg\max_{C_i} P(C_i|X) \\ &= \arg\max_{C_i} P(X|C_i)P(C_i) \\ &= \arg\max_{C_i} P(C_i)\prod_{j=1}^{n} P(X_j|C_i) \end{aligned} \quad (17)$$

The class with maximum probability leads to the suitable character for the input image.

VI. Experiments and Results

A. Experiments

In the experiments, for each region that represent an object from each color channel of the query image, the number of input features extracted using the order 3 of Legendre moments is 10 (L00, L01, L02, L03, L10, L11, L12, L20, L21, L30). The number of input features for color histogram is 16 per image channel. So, the result is 30 elements for Legendre moments and 48 elements for color histograms in the case of the 3 channels. The number of input features extracted using Texture extraction method is 14 x 6 = 84. These inputs are presented and feed to the Bayesian network classifier, for testing to do matching with the feature values in the reference database. To test the accuracy of the proposed approach, we measured the precision rates of each single descriptor.

Fig. 4 shows some examples of image objects from ETH-80 image database used in our experiments. The experiments are made based on different classes of objects.

Fig. 4. Some examples of objects from ETH-80 image database.

The accuracy of image annotation is evaluated by the precision rate which is the number of correct results divided by the number of all returned results.

All the experiments are conducted using the ETH-80 database containing a set of 8 different object images [30]. The proposed system has been implemented and tested on a core 2 Duo personnel computer using Matlab software.

B. Results

The results of the image annotation system based on Legendre, RGB, Texture descriptors and their fusion using a Bayesian network classifier are presented in table 1.

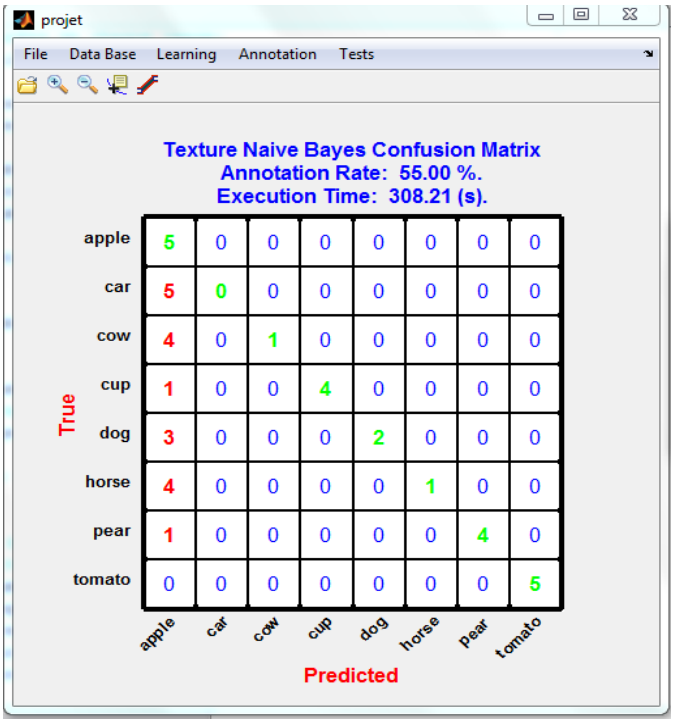
TABLE I. General Annotation Rates of the Annotation System Based on Legendre, RGB, Texture Descriptors and their Fusion using a Bayesian Network Classifier.

Descriptor Approch	Annotation rate (%)	Error rate (%)	Execution time (s)
Legendre	78.00%	22.00%	10756.49
RGB	67.50%	32.50%	162.16
Texture	55.00%	45.00%	489.65
Legendre+Texture+RGB+Bayes	87.50%	12.50%	12430.78

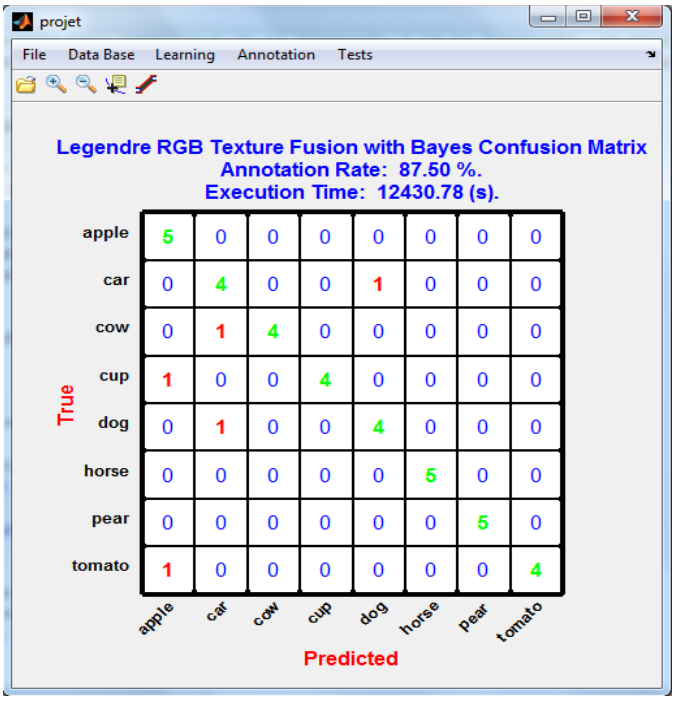
From the Table 1, the experimental results showed that the annotation rate of the proposed method based on the fusion of Texture, color histogram and shape descriptor increase the precision of the color image annotation system.

The fusion of descriptors will certainly increase significantly the annotation rate since they will fill each other. In some situation where one or two descriptors are not suitable, the 2nd or the other descriptors can give a good result. The object in the input image will be recognized and annotated by either the texture or the color histogram or the shape descriptors.

In order to show the robustness of the proposed approach, more details are provided by calculating the confusion matrix of some method as presented in Fig. 5.



a)



b)

Fig. 5. Confusion matrix for: a) Texture with bayesian network, b) Texture, color and shape descriptors fusion with bayesian network.

The 2 confusion matrix in Fig. 5 show that the misclassified and incorrectly annotated objects (indicated by red color) in the case of using texture as single descriptor and Bayesian network classifier are reduced in the case of using the proposed approach based on fusion of many descriptors kinds (Texture, color and shape).

The results are also affected by the accuracy of the image segmentation method. In most cases, it is very difficult to have an automatic ideal segmentation. Therefore, any annotation attempt must consider image segmentation as an important step, not only for automatic image annotation system, but also for the other systems which requires its use.

VII. Conclusion

In this paper, the approach based on a fusion of different descriptors is used for the automatic image annotation system. For this image annotation system, we discussed the effect of merging different type of descriptors. The texture, color histogram and shape descriptors are merged together in one feature vector and used to classify and annotate the input image by the suited keywords that are selected from the reference database image. The performance of the proposed method has been experimentally analyzed. The successful experimental results proved that the proposed image annotation system gives good results for image that are well and properly segmented. However, Image segmentation remains a challenge that needs more attention in order to increase precision and accuracy of the image annotation system. Also, the gap between the low-level features and the semantic content of an image must be reduced and considered for more accuracy of any image annotation system. Other segmentation method and other features extraction method must be considered for future work. The feedback of results can be investigated for the automatic image annotation system. Finally, the execution time must be decreased in order to use the online system.

References

[1] Datta, R., Joshi, D., Li, J., and Wang, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40, 2, Article 5 (April 2008), 60 pages DOI = 10.1145/1348246.1348248 http://doi.acm.org/10.1145/1348246.1348248.

[2] N. Vasconcelos and M. Kunt, Content-Based Retrieval from Image Databases: Current Solutions and Future Directions, Proc. Int'l Conf. Image Processing, 2001.

[3] Shao, W, Naghdy, G and Phung, SL, Automatic Image Annotation for Semantic Image Retrieval, Lecture Notes in Computer Science, 4781, 2007, 369-378. Copyright Springer-Verlag.

[4] Lei Ye, Philip Ogunbona and Jianqiang Wang, Image Content Annotation Based on Visual Features, Proceedings of the Eighth IEEE International Symposium on Multimedia (ISM'06), IEEE computer society, San Diego, USA, 11-13 December 2006.

[5] J. P. Fan, Y. Gao & H. Z. Luo. Integrating Concept Ontology and Multitask Learning to Achieve More Effective Classifier Training for Multilevel Image Annotation. IEEE Trans. Image Processing, vol. 17, no. 3, pages 407–426, 2008.

[6] X. J. Wang, L. Zhang, X. Li & W. Y. Ma. Annotating Images by Mining Image Search Results. IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 11, pages 1919–1932, 2008.

[7] R. C. F. Wong & C. H. C. Leung. Automatic Semantic Annotation of Real-World Web Images. IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 11, pages 1933–1944, 2008.

[8] J.Z Wang, Jia Li. Real-Time Computerized Annotation of Pictures, Pattern Analysis and Machine Intelligence, IEEE Transactions on Volume: 30, Issue: 6, pp. 985-1002, 2008

[9] Ballan, Lamberto; Bertini, Marco; Uricchio, Tiberio; Del Bimbo, Alberto, Social media annotation, 11th International Workshop on Content-Based Multimedia Indexing 2013 (CBMI), pp. 229-235, 17-19 June 2013.

[10] Qi Mao; Tsang, I.W.-H.; Shenghua Gao, "Objective-Guided Image Annotation," Image Processing, IEEE Transactions on , vol.22, no.4, pp.1585-1597, April 2013.

[11] Zhang, D., Islam, M. M., and Lu, G., A review on automatic image annotation techniques. Pattern Recognition, 45(1): pp. 346 – 362, 2012.

[12] Mustapha Oujaoura, Brahim Minaoui and Mohammed Fakir. Article: Image Annotation using Moments and Multilayer Neural Networks. IJCA Special Issue on Software Engineering, Databases and Expert Systems SEDEX (1):46-55, September 2012. Published by Foundation of Computer Science, New York, USA.

[13] Mustapha Oujaoura, Brahim Minaoui and Mohammed Fakir, Multilayer Neural Networks and Nearest Neighbor Classifier Performances for Image Annotation, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 11, pp.165–171, 2012. Published by The Science and Information Organization, New York, USA.

[14] Oujaoura, M.; Minaoui, B.; Fakir, M., "A semantic approach for automatic image annotation," Intelligent Systems: Theories and Applications (SITA), 2013 8th International Conference on , vol., no., pp.1–8, 8-9 May 2013, doi: 10.1109/SITA.2013.6560800.

[15] R. Sinan Tumen, M. Emre Acer and T. Metin Sezgin, Feature Extraction and Classifier Combination for Image-based Sketch Recognition, Eurographics Symposium on Sketch-Based Interfaces and Modeling, pp. 1–8, 2010.

[16] Frank Y. Shih, Shouxian Cheng, Automatic seeded region growing for color image segmentation, Image and Vision Computing 23, pp. 877–886, 2005.

[17] Aristidis Likas, Nikos A. Vlassis, and Jakob J. Verbeek. The global k-means clustering algorithm. Pattern Recognition, 36(2): pp. 451-461, 2003.

[18] Lior Rokach, Oded Maimon. Data Mining and Knowledge Discovery Handbook, Chapter 15: Clustering Methods. pp 321-352, Springer series, 2nd Edition, New York, October 1, 2010.

[19] Zijun Yang and C.-C. Jay Kuo, Survey on Image Content Analysis, Indexing, and Retrieval Techniques and Status Report of MPEG-7, Tamkang Journal of Science and Engineering, Vol. 2, No. 3, pp. 101-118, 1999.

[20] Ryszard S. Chora´s. Image Feature Extraction Techniques and Their Applications for CBIR and Biometrics Systems, International Journal Of Biology And Biomedical Engineering, Issue 1, Vol. 1, pp. 6-16, 2007.

[21] Chee-Way Chonga, P. Raveendranb and R. Mukundan, Translation and scale invariants of Legendre moments, Pattern Recognition 37, pp. 119 – 129, 2004.

[22] R. Haralick, K. Shanmugan, et I. Dinstein. Textural features for image classification. IEEE Transactions on SMC, 3(6) : pp. 610–621, 1973.

[23] Ann.Becker, Patrick Naim : les réseaux bayésiens : modèles graphiques de connaissance. Eyrolles.1999.

[24] J. Pearl, "Bayesian Networks" UCLA Cognitive Systems Laboratory, Technical Report (R-216), Revision I. In M. Arbib (Ed.), Handbook of Brain Theory and Neural Networks, MIT Press, 149-153, 1995.

[25] Sabine Barrat, Modèles graphiques probabilistes pour la reconnaissance de formes, thèse de l'université Nancy 2, Spécialité informatique, décembre 2009.

[26] George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers, the Eleventh Conference on Uncertainty in Artificial Intelligence, 1995.

[27] Philippe LERAY, Réseaux bayésiens : apprentissage et modélisation de systèmes complexes, Habilitation A Diriger Les Recherches, Spécialité Informatique, Automatique et Traitement du Signal, Université de Rouen, novembre 2006.

[28] Patrick Naïm, Pierre Henri Wuillemin, Philippe Leray, Olivier pourret, Anna becker, Réseaux bayésiens, Eyrolles, 3ème édition, Paris, 2008.

[29] Tom .Mitchell: Generative and discriminative classifier: Naïve bayes and logistic regression. Machine learning. Draft 2010.

[30] ETH-80 database image. [Online]. Available: http://www.d2.mpi-inf.mpg.de/Datasets/ETH80

Building an Artificial Idiotopic Immune Model Based on Artificial Neural Network Ideology

Hossam Meshref, Member IEEE
Computer Science Department
College of Computers and Information Technology
Taif University, Taif, Saudi Arabia

***Abstract*—In the literature, there were many research efforts that utilized the artificial immune networks to model their designed applications, but they were considerably complicated, and restricted to a few areas that such as computer security applications. The objective of this research is to introduce a new model for artificial immune networks that adopts features from other biological successful models to overcome its complexity such as the artificial neural networks. Common concepts between the two systems were investigated to design a simple, yet a robust, model of artificial immune networks. Three artificial neural networks learning models were available to choose from in the research design: supervised, unsupervised, and reinforcement learning models. However, it was found that the reinforcement model is the most suitable model. Research results examined network parameters, and appropriate relations between concentration ranges and their dependent parameters as well as the expected reward during network learning. In conclusion, it is recommended the use of the designed model by other researchers in different applications such as controlling robots in hazardous environment to save human lives as well as using it on image retrieval in general to help the police department identify suspects.**

Keywords—Artificial Immune Systems; Artificial Neural Networks

I. INTRODUCTION

There are many ways to describe what is meant by Artificial Neural Networks (ANNs). In essence, the artificial neural networks are modeled after the biological neural networks that constitute the basic building blocks of the nervous system. The biological neuron, which is the basic element of ANN, consists of: soma, axons, dendrites and the synapses. The biological neural network consists of many neurons connected together in a specific way to learn to perform a certain function. Therefore, to summarize, what is artificial neural networks? One of the well known original definitions is:"An artificial neural network is a massively parallel distributed processor that has a natural tendency for storing experimental knowledge and making it available for use," [1].

There are many functions that a human could need in his daily life. Starting from early childhood, that function could be learning how to hold a cup to drink and to adjust the hand eye coordination mechanism using specific neural networks to perform such task. As the network learns to perform a certain function, the whole experience is stored in the synaptic strength between neurons.

The focus of this research will be on single-layer and multi-layer artificial neural networks as a model to draw ideologies from. The theory and algorithms behind single-layer and multi-layer networks will be briefly discussed in the design section. Having defined Artificial Neural Networks, it is important to shed some light on the artificial immune system. According to Dasgupta and Nino, immunity could be regarded as: "The condition in which an organism can resist diseases, more specifically infectious diseases. However, a broader definition of immunity is the reaction to foreign substances, pathogens, which includes primary and secondary immune responses," [2].

Section two covers the literature related work. Section three describes different ANN structures as well as their general different training methods. Section four, on the other hand, focuses on the artificial immune networks and the interaction between antigens and antibodies. At section five, the new proposed artificial immune network structure is proposed. Section six has a discussion that offers analysis of the proposed ideologies. Finally, section seven wrap ups with the overall conclusion and future work.

II. BACKGROUND

Reviewing the literature during the past decade, it was found that, in general, there are many applications that benefit from the immune system's features like self non-self discrimination, specificity, and memory [3-7]. The main idea behind their research was having a network that adapts itself by adjusting the concentrations of its nodes. On the other hand, a few researchers focused on analyzing immune networks to utilize it as a complement to artificial neural networks to make Neural networks more effective [8]. On the other hand, other researchers focused on developing artificial immune networks as alternatives to artificial neural networks. Vertosick & Kelly combined the immune network theory with parallel distributed processing concepts to produce an alternative ideology to neural network architectures [9]. However, these research efforts, although effective, were done more than a decade ago. Recently, researchers worked on another line of research trying to use hybrid system of artificial immune systems and other systems, e.g. neural networks [10-12]. The goal of those hybrid systems was to get the best advantages of both systems, but the resultant systems were complicated.

Most of the aforementioned research made considerable efforts to highlight the importance of artificial neural networks as well as the artificial immune networks. However, integrating both network paradigms is not the goal of this research, and

neither is the replacement of artificial neural networks by a more robust artificial immune network. What is being approached in this research is an attempt in which the complexity of the artificial immune network is being perceived from an artificial neural network lens. It is an exploration of opportunities that is expected to lay a foundation of an array of research projects that will benefit from the findings of the proposed research. One major advantage of this research is to produce a simple, yet a robust, model for an artificial immune network that could be useful in many applications. Moreover, producing a simplified model of AIN is expected to encourage researchers in the field to deploy the designed model in their applications.

III. INVESTIGATE DIFFERENT ARTIFICIAL NETWORK PARADIGMS

A. Structure of Artificial Neural Networks

As mentioned earlier, Artificial Neural Networks build upon the model of the biological nervous systems. Based on the information about the function of the brain within the nervous system efforts were invested to obtain a mathematical model for the human learning habits. The results of these efforts laid the foundation of Artificial Neural Networks [13]. The mathematical models started from the smallest entity in the nervous system: the neuron. Scientists started with introducing the mathematical model of the artificial neuron: The Perceptron.

A Perceptron has a set of n synapses associated to the inputs. Each of them is characterized by a weight. Each input signal, x_i is multiplied by its corresponding weight, w_i. The weighted input signals are summed and a bias, $x_0 w_0$, is added to improve learning. Thus, a linear combination of n input signals is obtained, see equations 1 and 2. A nonlinear activation function φ is applied to the weighted sum Z, and the final output is expected to fire only if it exceeds a threshold θ.

$$Z = \sum_{i=1}^{n} x_i w_i + x_0 w_0 \qquad \text{eqn. (1)}$$

$$f(z) = \begin{cases} 0, & Z < \theta \\ 1, & Z \geq \theta \end{cases} \qquad \text{eqn. (2)}$$

The activation function could have many forms according to the neuron design (e.g. Log-Sigmoid function and Tan-Sigmoid function). In the case of the Perceptron, the design is simple and entails using the threshold function as it is required to classify two-class inputs only [14]. The Perceptron was followed by the introduction of single-layer and multiple-layers neural networks. In single-layer ANN the input layer acts as a fan-in layer that does no processing. However, in the output layer, each node is a neuron, which receives inputs and produces an output according to the previously mentioned process of the Perceptron. The activation functions used at the output layer may be any of the aforementioned functions and not necessarily the threshold function. Multi-layer ANN, on the other hand, consists of more than one processing layer. Considering feed-forward ANN, all calculations are done in parallel in each layer, and the output of one layer is broadcasted to successive layers until the final network outputs are calculated. Only the last layer is called the output layer, while all the layers located between the input and output later are called hidden layers, see figure 1.

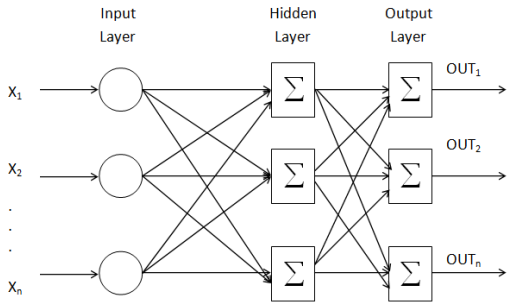


Fig. 1. Multi Layer Neural Networks

B. Training of Artificial Neural Networks

In the previous section, artificial neural networks were categorized according to their structure. Artificial Neural Networks could be also categorized according to their method of learning. Its ability to learn resembles the method that the human brain learns. In general, researchers categorized the learning methods of artificial neural networks to be within three main categories: Supervised Learning, Unsupervised Learning, and Reinforcement Learning [15].

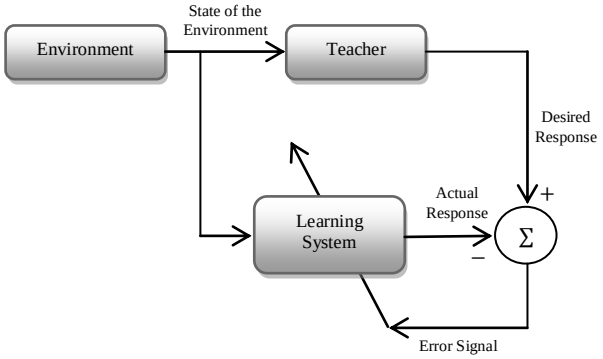


Fig. 2. Supervised Learning of Artificial Neural Networks

During training using supervised learning, the network learns and gains experience from a set of predefined training examples, see figure 2. During the training session the input vectors are applied to the network, and the resulting output vector is compared with the desired response. If the actual response differs from the target response, an error signal adjusts the network weights. The error minimization process is supervised by a teacher. In general, supervised training method is used to perform non-linear mapping in pattern classification nets, pattern association nets, and multilayer nets.

On the other hand, there is another method of training: the unsupervised learning method. This type of training doesn't require a teacher. In this method, input vectors of similar types are grouped together without a teacher. After training, when a new input pattern is applied, the NN provides an output response indicating the class to which the input pattern belongs. If a class cannot be found, a new class is generated.

The third method of training artificial neural networks is the reinforcement training method, see figure 3. In that method the network learns by trial and error, and the performance of each element is being monitored, which means that there is a certain level of supervision, but there is no predefined desired output. Instead, there is a common goal for the whole network to achieve. Network elements perceive their states and perform actions. After each round of actions, the performance of each element is evaluated and the corresponding critic is assigned. This process is repeated until the network overall goal is achieved.

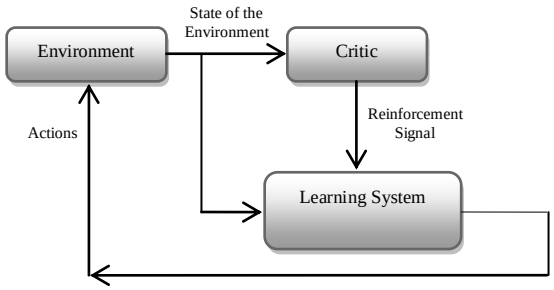


Fig. 3. Reinforcement Training of Artificial Neural Networks

IV. Investigate the Artificial Immune Network Paradigm

The biological immune system consists of lymphocytes that have two major types, T-cells and B-cells. B-cells are responsible for humoral immunity that secretes antibodies. T-cells are responsible for cell mediated immunity. Each B-Cell has a unique structure that produces suitable antibodies in response to invaders of the system. That type of response is called innate immunity and eventually results in antibody-antigen relations to be stored in case the host encounters the same invader again. In that case, the immune response in expected to be faster given that the network has seen it before, i.e. learned how to deal with it [16, 17].

The immune system has Idiotopic networks that use stimulation and suppression among its elements to achieve immunity against antigens (Ags). The part of an antigen that can be recognized by antibodies is called epitope (Ep). As a result of this stimulation the B-cells start to produce Abs (Y-shaped). On the other hand, the part of the antibody that can recognize epitopes of antigens is called paratope (P). However, part of an antibody, called idiotope (Id), could be regarded as antigen by other antibodies in the idiotopic network [18, 19], see figure 4.

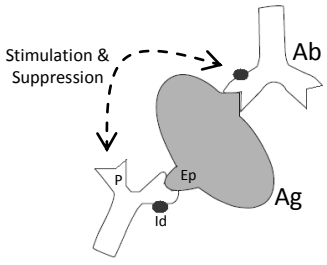


Fig. 4. Detailed Stimulation and supression in Idiotopic Networks

V. Building the New Idiotopic Immune Network Based on the Most Suitable Artifcial Neural Network Model

In the previously reviewed literatures some researchers initiated an idea about how the AIN could be viewed from an ANN ideology. In their depiction, the network structure was abstract, while the pool of Idiotopes and Epitopes was loosely connected. However, there is a need for a more precise realization in order to have a robust artificial immune Network. It is our goal in this research to clarify the depiction of artificial immune networks from the artificial neural networks point of view by giving an explanation for the investigated ideology. That could be done by starting the design from the original constructing Lymphocyte units, and that will eventually lead to a concrete Artificial Immune Network structure.

A. Network Structure Primary Design

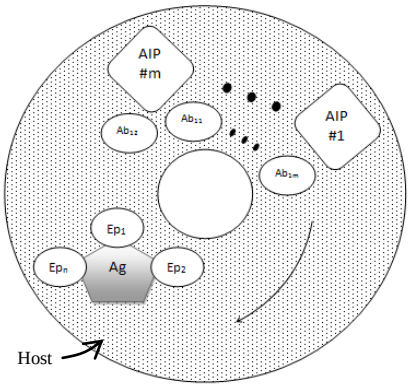


Fig. 5. Stage-1 Ag stimulation

The artificial immune network is not necessarily constructed out of layers as the case in Artificial Neural Networks. Therefore, this project design started from the basic Perceptron unit, B-Cell, and builds its way up to the full multi Artificial Immune Perceptron (AIP) network.

Figure 5 illustrates stage-1 when an Ag attacking a host. As the features of that Ag, Eps, appear in the host, matching Abs are being produced by different AIPs when being stimulated by the Ag. Based on the concept of complimentarity, each Ep is paired with its corresponding Ab Paratop P, and eventually a set of matching pairs *(Ep, P)* are produced. For example, AIP#1 could produce antibodies Ab_{11} and Ab_{12}. Those antibodies could bond with the Ag Epitop Ep_1 if they match it. That process is called affinity measurement. The Ab required concentrations are then decided, affinity maturation, and then sent back to the host to counter the effect of the Ag. Figure 6 illustrates stage-2 where Abs are secreted by the AIPs to suppress Ag stimulation.

B. The Final Design of the Multi-AIP network

The stages depicted in figure 5 and 6 could be combined to form one network structure that resembles the structure of a single layer artificial neural Perceptron network, see figure 7. The input layer has the Ag features Ep_1, Ep_2... Ep_n. These features are passed to the output layer where affinity measurement and affinity maturation are handled.

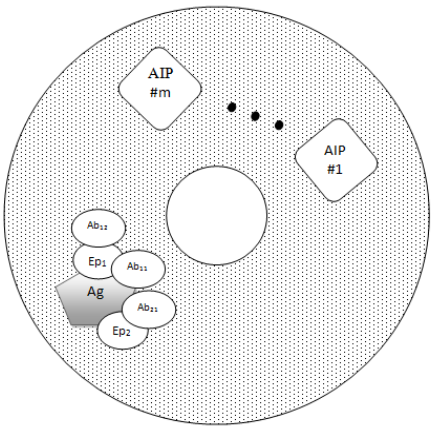


Fig. 6. Stage-2 Ab supression

The generated Ab's are fed back to the first layer to counter the effect of the Ag Ep's.

Based on the nature within the network, Farmer et al claimed that the interaction within the Idiotopic immune network, which is modeled in the Multi-AIP network, could be governed by the following equation, equation 3:

$$\frac{dC_i(t)}{dt} = (\sum_{j=1}^{N} m_{j,i}c_j - \sum_{k=1}^{N} m_{i,k}c_k + m_i - k_i)c_i(t) \quad \text{eqn. (3)}$$

,where the first term, from the left, represents the total stimulation between different B-cells in the idiotopic network. The second term represents the total suppression. The third term represents the Ab stimulation received by an Ag based on the affinity between them. As for the fourth term, *k*, it represents the mortality of an Ab [20]. *N* represents the number of Abs, while *C* represents the concentration of Ab within the network, and *m* represents affinity.

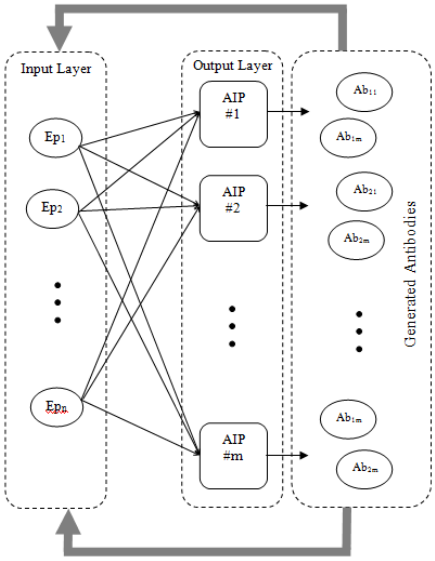


Fig. 7. The Multi-AIP Network Structure

VI. Discussion

The dynamics of the aforementioned Multi-AIP network structure could be described by explaining the dynamics of its components: antigens and antibodies. However, not all the artificial neural network models could be deployed in this artificial immune network designed model. For example, an artificial neural network models that are based on supervised learning are not likely to be implemented. The reason is that the concept of forward pass and reverse pass would not be fully implemented as a technique for handling the network dynamics. In particular, the concept behind the reverse pass dynamics cannot be deployed, because there is no desired output to compare with the actual output produced at the output layer of the network. However, it was found that there is a common feature among most artificial neural network models. Most models store the learned experience in the weights associated with each input, and in general these weights are updated during the learning process according to equation 4:

$$w_{t+1} = w_t + \Delta w \quad \text{eqn. (4)}$$

This equation means that the new weight during learning at time *t+1* is based on the previous value of weight at time *t* plus an increment Δ*w*. The value of this increment varies according to the learning method. By the end of the learning process the learned knowledge is expected to be stored in the weight values that represent the strengths between network components. Similarly, in the artificial immune network antibodies concentration values changes during learning to suppress antigen behavior. In the end, the learned knowledge is expected to be stored in the concentration values. By projecting the weight update concept on the artificial immune network learning process, it could be concluded that antibody concentration could follow the same ideology:

$$C_{t+1} = C_t + \Delta C \quad \text{eqn. (5)}$$

To accurately model concentrations for antigens and antibodies, theoretical biology should be involved. According to Boer, population growth could be described by a classical logistic equation. Therefore, in this research, antibody concentration could be modeled using equation 6.

$$C_{Ab}(t) = \frac{C_{Ab_\circ} + C_{Ab_{max}}}{C_{Ab_\circ} + (C_{Ab_{max}} - C_{Ab_\circ})e^{-rt}} \quad \text{eqn. (6)}$$

$$r = b - d$$

, Where $C_{Ab_\circ}$ is the initial concentration of the Ab population, $C_{Ab_{max}}$ is the carrying capacity of the population, and r is the natural rate of increase represented as the difference between the birth rate, b, and the death rate d. Therefore, by controlling $C_{Ab_{max}}$ and $C_{Ab_\circ}$ the Ab population size could be controlled. However, it was noticed that as *r* increases, the concentration curve becomes more steep (i.e. when $r_2 > r_1$), see figure 8.

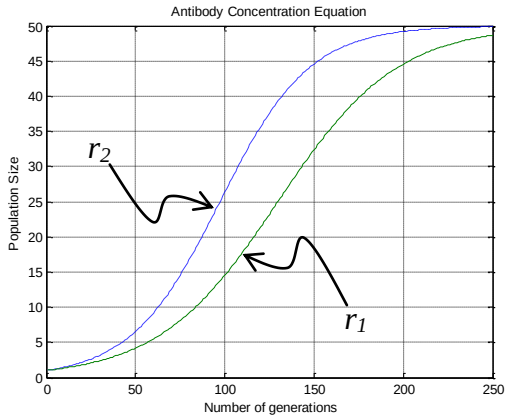


Fig. 8. Antibody Concentration

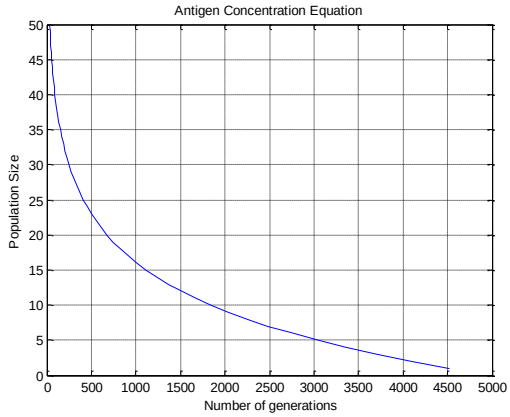


Fig. 9. Antigen Concentration

The Ag population mathematical representation is in general similar to the Ab abovementioned representation. However, in this research, the Ag is not assumed to proliferate, *b=0*, which denotes that the Ag population will ultimately extinct and its concentration will eventually reach zero, see figure 9. Unlike the Ab concentration equation, antigen concentration could be modeled using equation 7:

$$C_{Ag}(t) = C_{Ag_{max}} e^{-dt} \quad \text{eqn. (7)}$$

As learning progresses, Ab and Ag concentrations are expected to change. The value of ΔC determines the amount of expected change that takes place during that process. However, as mentioned earlier, that value varies as the learning mechanism changes.

Based on the nature of the Multi-AIP network structure, and the expected interaction between Paratop-Epitop as well as Paratop-Idiotop, the best ANNs model to use would be the one that is based on the competitive network ideology. Therefore, according to reinforcement learning ΔC could take the following values during learning as depicted in table I:

TABLE I. ΔC VALUES DURING A LEARNING EPISODE

Time Step	Value of ΔC
Step#1	μ_t
Step#2	$\mu\, r_{t+1}$
Step#3	$\mu^2\, r_{t+2}$
...	...
Step#n	$\mu^n\, r_{t+n}$

, where μ is a discount fraction, between 0 and 1, raised to the power of the number of time steps, and the reward at time *t* is r_t. More succinctly, the change in Ag and Ab concentration could be given by the following equation:

$$\Delta C_t(n) = \mu^n r_t \quad \text{eqn. (8)}$$

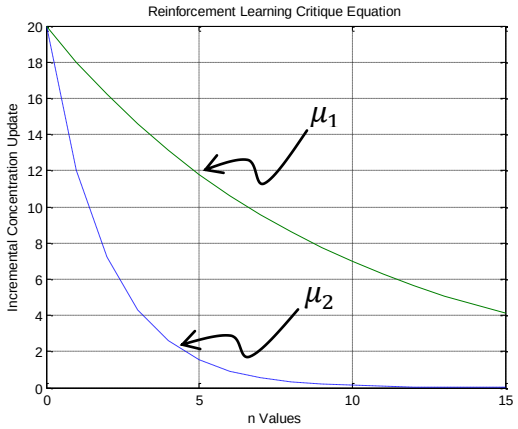


Fig. 10. Antibody Concentration

The result of the analysis showed that as the number of iteration steps increases, the incremental change an Ag and Ab concentrations also increases see figure 10. However, in general it is concluded that as $n \to \infty$ the value of $\Delta C_t(n) \to 0$. In addition, results showed that as the value of the discount factor μ decreases, the curve becomes less steep ($\mu_1 < \mu_2$).

VII. CONCLUSION AND FUTURE RECOMMENDATIONS

Artificial neural network has been a reliable model for at least the past two decades. In this project, it proved that it can deliver ideologies to inspire the creation of a robust artificial immune network model. There were many artificial neural network models in the literature, but current research analysis showed that the reinforcement learning model suits the dynamics of the artificial immune networks. Despite the diversity of modeling techniques for artificial neural networks, careful selection has been deployed to store the network learned experience as well as the way to update it. Finally, the effect of some network learning parameters values were investigated to better control the network learning process. That resulted in a more robust artificial immune network.

In conclusion, it is believed that the designed model could be used for many applications.

The careful design scaffolding process that was implemented in building the model for an artificial immune network led to a robust model. The model in general suits different applications that deploy intelligent behavior. For future work, it is recommended to use the designed network model to build some applications in the Robotics field, for example, where robots can be used in rescue operations. These types of applications can intervene in hazardous situations and save many human lives. Another area that is recommended could be in pattern recognition to help the police department identify suspects even if partial interrogation information is available. That can narrow down the suspects' list, and eventually save time, money, and particularly huge office work efforts. These are some examples of applications that could benefit from the proposed designed model, but in general, there are several applications especially in the artificial intelligence field that could benefit from it.

REFERENCES

[1] Simon Hykins, Neural Networks: A Comprehensive Foundation, Prentice Hall, 1999.

[2] Dipankar Dasgupta, Luis F. Niño, Immunological Computation: Theory and application. NewYork, NY : CRC Press, 2009.

[3] Purbasari, A., Iping, S.S., Santoso, O.S., Mandala, R., "Designing Artificial Immune System Based on Clonal Selection: Using Agent-Based Modeling Approach," *Modelling Symposium (AMS), Asia*, pp.11,15, 23-25 July 2013.

[4] Chung-Ming Ou, Yao-Tien Wang, Ou, C.R., "Intrusion detection systems adapted from agent-based artificial immune systems," *Fuzzy Systems (FUZZ), 2011 IEEE International Conference* , pp.115,122, 27-30, June 2011.

[5] D. Dasgupta, "Advances in Artificial Immune Systems," IEEE Computational Intelligence Magazine, Vol. 1, No. 4, pp. 40-49, 2006.

[6] Meshref, H., Vanlandingham, H., "Artificial Immune Systems: Application to Autonomous Agents," *IEEE International Conference on Systems, Man, and Cybernetics*, Nashville, Tennessee, Vol.1, pp. 61 – 66, 2000.

[7] Meshref, H., Vanlandingham, H., "Immune network simulation of reactive control of a robot arm manipulator," *Soft Computing in Industrial Applications, SMCia/01*, Proceedings of the 2001 IEEE Mountain Workshop on Soft Computing in Industrial Applications, pp.81,85, 2001.

[8] Seral Şahan, Salih Güneş, "Immuno Theory Applications in Neural Networks," retrieved from The Chamber of Electrical Engineers (EMO) at http://www.emo.org.tr/ekler/19345a4c56c55ba_ek.pdf on July 2012.

[9] Vertosick, F. T., Kelly, R. H., "The Immune System as a Neural Network: A Multi-epitope Approach," Journal of Theoritical Biology, pp.225-237,1991.

[10] Seral Şahan, Kemal Polat, Halife Kodaz, Salih Günes, "A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis," Computers in Biology and Medicine, Vol. 37, No. 3, pp. 415-423, 2007.

[11] Ruiying Zhou, Qiuhong Fan, Mingjun Wei, "Solving for multimodal function with high dimensions base on Hopfield Neural Network and immune algorithm," *International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT)*, China, August 2011.

[12] Huang Yue , Shenyang Ligong, Li Dan , Gao Liqun, " Power system short-term load forecasting based on neural network with artificial immune algorithm," *24th Chinese Control and Decision Conference (CCDC)*, Taiyuan,China, May 2012.

[13] Kishan Mehrotra, Chilukuri Mohan, Sanjay Ranka, Elements of Artificial Neural Networks, Bradford Press, 1996.

[14] Colin Tosh, Graeme Ruxton, Modelling Perception with Artificial Neural Networks, Cambridge University Press, 2010.

[15] B. Yegnanarayana, Artificial Neural Networks, Prentice Hall of Indea, 2006.

[16] De Castro, L. N., J. Timmis, Artificial Immune Systems: A New Computational Intelligence Approach, Springer, Heidelberg, 2002.

[17] Dipanker Dasgupta, "An Overview of Artificial Immune Systems and Their Applications," Springer, 1998.

[18] N. K. Jerne, "The generative grammar of the immune system," The EMBO Journal, Vol.4, No.4, pp. 847–852, 1985.

[19] N. K. Jerne, "Idiotopic Network and Other preconceived ideas," Immunological Rev., Vol. 79, pp. 5–24, 1984.

[20] Farmer, J. D., N. H. Packard, A. S. Perelson, "The immune system, adaptation and machine learning," Physica, 22D, 187–204, 1986.

Anonymous Broadcast Messages

Dragan Lazic
School of Computer Science
University of Guelph
Guelph, ON, CANADA

Charlie Obimbo
School of Computer Science
University of Guelph
Guelph, ON, CANADA

***Abstract*— The Dining Cryptographer network (or DC-net) is a privacy preserving communication protocol devised by David Chaum for anonymous message publication. A very attractive feature of DC-nets is the strength of its security, which is inherent in the protocol and is not dependent on other schemes, like encryption. Unfortunately the DC-net protocol has a level of complexity that causes it to suffer from exceptional communication overhead and implementation difficulty that precludes its use in many real-world use-cases. We have designed and created a DC-net implementation that uses a pure client-server model, which successfully avoids much of the complexity inherent in the DC-net protocol. We describe the theory of DC-nets and our pure client-server implementation, as well as the compromises that were made to reduce the protocol's level of complexity. Discussion centers around the details of our implementation of DC-net.**

Keywords—Dining Cryptographer network; Privacy; sender-untraceability

I. INTRODUCTION

The issue of privacy and anonymity on the Internet has become a challenging one, especially with the growing influence that the Internet has on our day-to-day social lives. With the increased use of social networking sites and mobile Internet based communication, being anonymous and maintaining privacy has becoming something that the average computer has great difficulty in achieving. Preserving the anonymity of the communicating parties is crucial in situations where knowledge of the identity of the source of communicated messages could create a conflict of interest, jeopardize the integrity of a process or endanger the participants. The concept of maintaining anonymity is simple to conceive. However, they can be rather difficult to implement. The trails left by the protocols and technologies involved in digital communication can be difficult to erase or hide to a point where the identity of the communicating parties is not exposed. The trails created in these communications are often required for the communication itself often to maintain a quality of service or integrity of the message being communicated. This paper describes a digital, computer based form of communication that preserves the anonymity of all communicating parties. The program takes heavy influence from David Chaum's Dining Cryptographers problem and the DC Net concept.

A. Background

According to Nissenbaum [1], online anonymity is "unreachability", i.e. the inability of communication, or action of an individual to be traced to a specific person at a specific address in the real world.

A precise, mathematical definition of anonymity has been elusive [2][3]. Uncommon attributes of an individual may still be used to "fish-out" a person's identity, even though one may not know their name, phone number, or address explicitly from a database [4][5][6][7].

Online anonymity has its pros and cons in the society. It can provide a means for free speech and criticism of established power without fear of reprisal [8]. An example of this is when it was used in the media of communication in the North Africa, Middle-East uprising [9][10][11]. The essence of anonymity – and the need to assure deterrence from repercussions created the need for the setting up of a system that protected message conveyors from being identified in the quasi-changed political systems. However, online anonymity can also have detrimental effects in the society. Examples of this include anonymous hacking [12], and communication by terrorists [13].

The anonymous communication method described in this paper is based upon the Dining Cryptographers problem. The Dining Cryptographers problem was first proposed by David Chaum [14][15] in 1988. David describes a thought experiment and proposes a solution, which he develops into a theoretical Dining Cryptographers Protocol (AKA. DC-net) that can be used for broadcasting of unconditional anonymous messages (Chaum 1988). Prior to the development of the DC-net protocol, Chaum developed the concept of multiparty-secure sender-untraceability protocol' which he called the mixed-net protocol. The mixed-net protocol idea was then used and actually implemented in the onion routing protocol of TOR.

To illustrate the theory behind the DC-net the story of the dining cryptographers is often used. The original Dining Cryptographers problem begins with three cryptographers having dinner together at a restaurant. Their waiter informs them that arrangements had been made for the bill to be paid anonymously. One of the three cryptographers might be paying for the bill, or it could be the U.S. National Security Agency (NSA). The three cryptographers want to respect each other's right to privacy but they would also like to know if the NSA covered the bill. They devise their plan, while hiding behind their menus they each flip a coin so that only the person sitting on their right can see. Then they say aloud if the face of the coin they flipped coincides with the face of the coin flipped by the person to their left.

An odd number of differences uttered by the cryptographers would indicate that one of them paid (assuming that the bill was paid once). Yet the payer remains anonymous to the rest of the Diners in that case.

For simplicity, the truth table below indicates the results of the sums of the differences uttered as a result of the comparison of the coin-toss, had none of them paid. It is easy to verify that the above statement is correct.

TABLE I. RESULTS OF THE SUMS OF THE DIFFERENCES UTTERED AS A RESULT OF THE COMPARISON OF THE COIN-TOSS

D1	D2	D3	Differences of Utterances
T	T	T	O
T	T	H	2
T	H	T	2
T	H	H	2
H	T	T	2
H	T	H	2
H	H	T	2
H	H	H	0

1) Peer-to-Peer

With peer-to-peer the clients all make connections to each other in a ring design. For each bit that is transmitted, the results from the first stage of the DC-net protocol are combined by sending them around the ring network. Each user sends their result to the "next" user in a previous determined direction on the ring network. Now each user has 2 stage-one result bits and XORs them. Then they pass the XOR result onto the next user. This process repeats until the results have Figure 1: Dinning Cryptographers

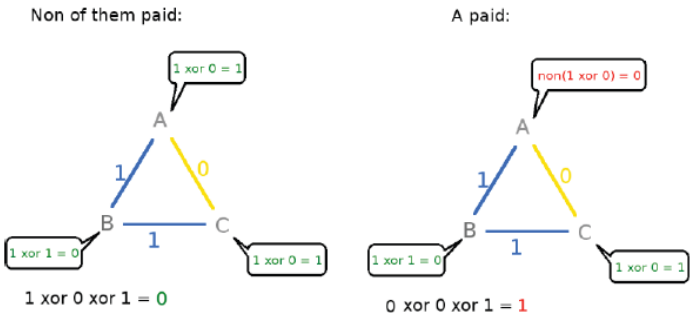


Fig. 1. Dinning Cryptographers

made an entire circuit of the ring (equal in length to the number of participants). Now every member of the ring should have the stage-two result and the entire process repeats for the next bit. The obvious problem with this implementation the overhead of waiting for each client to process then send and also the restructuring that would have to happen every time a new client joins the ring. Figure 2 shows a basic layout of Peer-to-Peer connection.

2) Hybrid Client-server

In this implementation the users send their results to a server, which XORs them and sends back the results. However the pair-wise shared secrets between clients is still communicated in a peer-to-peer manner with direct connections between peers.

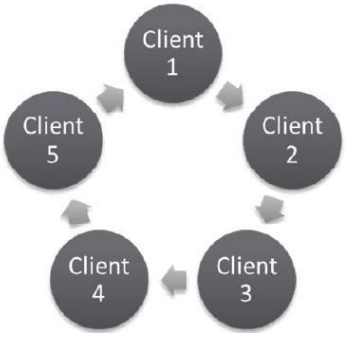


Fig. 2. Peer-to-Peer Model

The hybrid client-server implementation has the advantage of communication efficiency compared to the peer-to-peer implementation. The communication overhead for each client is drastically reduced since a circuit of the ring network doesn't need to be made in order to broadcast messages. The primary disadvantage of this implementation is the necessity of a server in addition to the existing peer-to-peer connections, which increases complexity. Figure 3 shows a basic layout of the Hybrid Client-server connection.

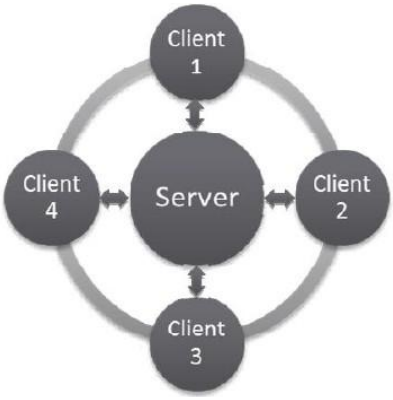


Fig. 3. Hybrid Client Server Model

II. LIMITATIONS

The DC-net protocol is straightforward and elegant in theory. It's also highly secure. However, it has several limitations, which may make it undesirable in certain use-cases. The limitations of the DC-net protocol can be generally grouped into three areas: collisions, disruption and complexity.

A. Collisions

Due to the nature of the DC-net protocol only one byte can be processed at a time. Otherwise if multiple clients send a message the XORed result on the server at the end would be unreadable text. To mitigate this problem, a system of start and end messages will be used. Before a clients message is sent to a server, a specially selected start character is sent before the message. This start character means that the server will then know to expect a message and that the end of the message will be followed with an end-message character. During that time it send a message to the clients notifying them that a message is being received and to not send.

Another similar collision issue is a race condition for sending the start message. To address this a check was put in place, if the XORed result on the server side came out to be anything other then a 0 or a start message the server would then not print the message but instead wait until it was XORed zeroes again and then sent a broadcast notifying of a collision and asking the clients to send again.

B. Complexity

The level of complexity and overhead that exists and can be added to this protocol is rather larger. The existing complexity adds to the communication overhead, which only gets worse with more connections. The clients are constantly sending data to the server, which then has to be processed by the server, even when no actual message is being broadcast. Also if you were to choose to encrypt the connection that would add additional complexity and overhead for both the client and server and decrease performance further. All these overhead results decrease the performance of the server, and as more clients connect performance continues to drop. However this is a necessary limitation and without it, their would be no anonymity.

C. Integrity

The DC-net protocol has no way of checking the integrity of the clients or the server. Essentially this allows anyone to do as they please within certain confines. For example a rouge server can be hosted that works at identifying the clients that connect and broadcast, or a client might be capable of jamming someone or the server from broadcasting. Adding additional checks in place to prevent this kind of action however could result in the loss or the risk of loosing anonymity.

III. Implementation

The implementation deed 24asone in this project is different from the ones outlined earlier [16]. The Client-Server scenario detailed in the previous section required too much communication overhead [17], and the peer-to-peer scenario was even worse. In the archetypical DC-net protocol model discussed previously, both stages of the DC-net protocol are performed for each single bit of data transmitted. This was determined to be extremely wasteful. To lower communication overhead of the DC-net protocol, random number generators were introduced to replace the function of "coin flips". This allows DC-net protocol rounds to be conducted 8-bits at a time.

The server is implemented as a dedicated application that acts as a broadcast hub for the clients as well as the calculator for stage two of the DC-net protocol. The server cannot participate in the chat in the same manner as a client would, however it does send informational broadcasts when required. However, just like in the original Client-Server DC-net scenario, the server is used to keep sessions of the chat. The clients have the ability to send and receive messages. The clients handle stage one of the DC-net protocol and send the results to the server. The clients have two operating rooms, one for regular chat and another for anonymous chat. When all clients have entered into the anonymous room and indicated their readiness the anonymous mode will begin and a count down timer for their session will also start.

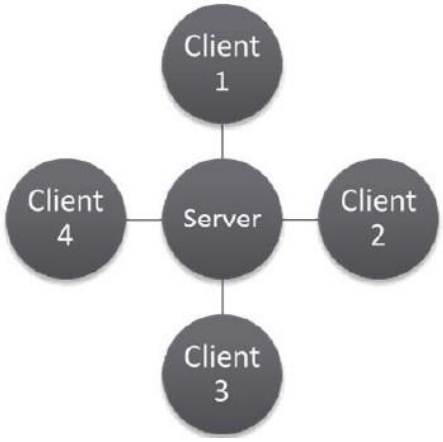


Fig. 4. The client-server connection model

A. DC-Net Implementation

Given the two-sided client-server design and the two stages involved in the DC-net, implementation was done in two stages. The first stage was to ensure that there was an operating general chat program to work of off. The bases for the chat program was found online and modification were made to insure it better fit the needs of the project [18]. The general chat program offers a simple chat interface that requires a unique nickname and the IP address of the server. Once the chat client was done adding the DC-net protocol was the next stage and the bulk of the implementation.

Once the client connects and enters the anonymous room and everyone is ready, the server generates a random seed for each client in the ring at the point of connection. This random seed is then sent to each client as well as the client to their "left". Therefore each client will have two seeds. Once anonymous mode has been activated in the room, the client uses the two seeds to generate the results of the 'coin toss'. The DC-net protocol cycle operates thusly. During stage-one, each client uses their seeded random number generators to produce two sets of bits, 1 byte each in size. These two sets are XORed to produce a single stage-one-result byte. The clients then immediately send the stage-one-result byte to the server using the primary communication channel (implemented by a TCP byte stream).

The server logs the stage-one results as they are received but performs no further action until all stage-one result bytes have been received from every client. Once all the results have been logged from every participating client, the server XORs all the results in the log, producing the stage-two result. Finally the server broadcasts the stage-two result to all clients using the secondary communication channel (implemented by TCP text stream). When the clients receive the stage-two result, a new round of the DC-net protocol is triggered.

B. Client Classes

The client is a stand-alone program consisting of several classes.

1) ClientRunner.java and ChatClient.java

ClientRunner is a very simple and small class. It essentially is used to start the client. ChatClient, is the class containing all

of the code for the GUI. This class also handles creating and executing the socket connection for the basic chat client. Besides the methods used for the creation and handling of the GUI, the class also has two other methods, one that creates a byte stream over the socket connection to receive data and another that does the same except to send data. Figure 5 shows the GUI design for the normal chat room while Figure 6 shows the design for the anonymous chat room.

2) ServerReader.java

The class starts an infinite loop and, using the receiver method created in the ChatClient class, it listens on the read byte stream for any data sent in by the server. Once the class finds data on the stream, it will act in accordance to that data.

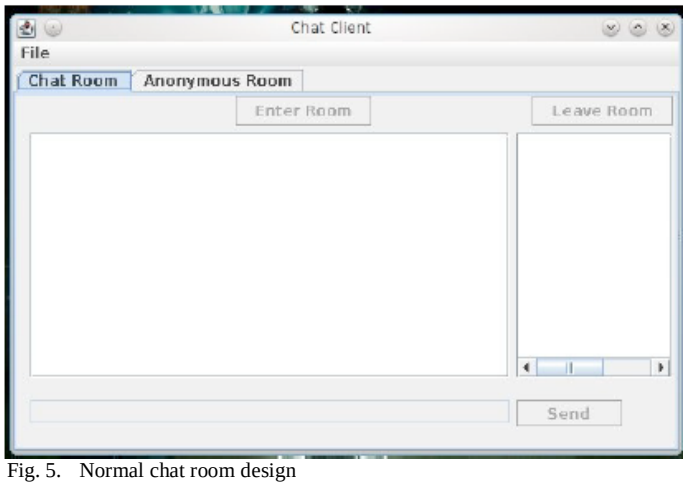


Fig. 5. Normal chat room design

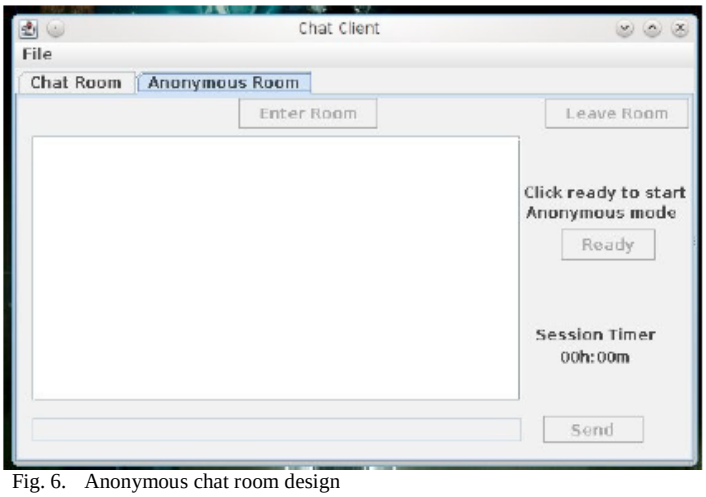


Fig. 6. Anonymous chat room design

This class essentially acts as a communication hub for the client to the server. Should the server need to send anything relevant to the client it will be received and handled by this class. If the server needs to send notice of a collision or that a message is being received then this class will receive the message and print it in the client chat window.

3) ByteSender.java

When anonymous mode is activated this class is triggered and starts an infinite loop. It uses the seed given to the client by the server and also by the other client, to generate the pseudo-random 'coin tosses'. These coin tosses will be represented by a string of numbers. Each number's binary value will equal 8 coin tosses. Once the coin tosses have been generated they are XORed. The class then makes a check against a Boolean variable to see if the user has made an inputted a message in the chat client. If the user has inputted a message it will start reading the bytes of the message and XORing them against the results of the XORed generated coin tosses. If the user doesn't input a message then only the XORed results of the coin tossed will be sent to the server. This procedure continues until the client receives a signal to stop

C. Server Clases

1) Server.java

The Server class creates many of the foundation objects for the DC-net protocol and for the chat environment in general. The class creates the vectors collection that contains the socket connection for each client, the socket objects for both the client and the server and the generator for the pseudorandom seed to be used by the client.

The class starts an on-going loop as it waits for an incoming connection from the clients. Once a connection is established it creates a user land thread for the client that made the connection. The thread is then added to the appropriate vector object in the collection. The class then creates an instance of another class, CThread, and assigns it to the recently added client thread. The Server class continues this process for every client connection it makes.

2) CThread.java

The CThread class creates the required objects and posses the needed methods to cover all the actions executed by the client. It begins by calling the input and output streams created as a result of the clients connection to the server. It then starts an infinite loop, checking the IO streams for traffic and reacting appropriately to any traffic. The class also acts as the main point of communication between the client and the server on the server side.

All messages sent by the client, whether they are meant for the server or for broadcast will come down to and be handled by CThread. The class will also handle any operations that the server must make due to the client. For example, if the client moves rooms from the general chat to the anonymous room or if they client gets put on a wait list. The class will check if the clients nickname is unique, and if it meets the requirements to enter anonymous mode. If the client cannot enter anonymous mode CThread will notify the client why. Overall the CThread class handles a lot of the smaller tasks related to the Client-Server interaction.

3) ByteReader.java/SeedSender.java/SessionTimer.java

SeedSender, will parse through the vectors collection and send each active client object in the vector their appropriate seeds. SeedSender will also notify the clients that the server is ready once all the seeds have been distributed. Afterwards SeedSender will call the ByteReader class and terminate itself. Once called, ByteReader will immediately call the SessionTimer class to start tracking the session time. An infinite loop is then created, and the class calls a method in CThread of each client to fetch a byte of data coming in from

the client. This will essentially be the XORed results that the clients are continually sending to the server. The fetched byte will be stored in a byte array, using one element for each client. It then takes the bytes and XORs them together as it should for a stage two process. It will also convert the XORed result to a string and check the character result. If the result is a start message then the server will know a message is about to be sent. Otherwise it should result in a zero if there was no collision.

When SessionTimer is called it will set the timer for every active client to 15min. This will be the default duration of the anonymous session. As it counts down, at every minute interval it will update the timer for each client. Once the timer has hit the 0 mark, it will check the waiting list for the anonymous room and move over any clients on the waiting list into the room and reset itself.

IV. RESULTS

The implementation of the DC-net protocol created for this project is a functional implementation. It operates in as described by the DC-net protocol and performs both stages of the DC-net protocol for each bit of data sent. However it is not a complete or direct implementation of the protocol as it eliminates all the peer-to-peer communications. The implementation is also relatively practical. It offers both standard user-attributable messaging as well as anonymous messaging. By eliminating all peer-to-peer communication and centralizing the entire DC-net protocol to operate through a server, I have also significantly reduced the communication overhead associated with the DC-net protocol. The pure client-server implementation developed for this project also largely solves the problem of complexity in the DC-net protocol. With all communications centralized at the server, there is reduced overhead and a reasonable number of network connections.

A. Known Issues

The program is not perfect and therefore has some unresolved bugs.

- Leaving the room while waiting for anonymous mode to initiate will not be recognized by the server
- Server notice of incoming message will be attached to a client message if connection speeds are very fast. Example: running on a Local host or a fast LAN.

This is just a list of what was found in test.

V. CONCLUSION

There are cases where privacy preservation is important even in instances where communication is taking place. Examples of this include online-surveys. The Dining Cryptographer network (DC net) was devised by David Chaum for anonymous message publication. This is an elegant and straightforward protocol in theory.

However, it has three main drawbacks, Collisions, Complexity and Integrity. Collisions: according to the protocol, only one byte can be processed at a time, multiple clients sending a message would result in a futile attempt at XORing.

A. *The DC net is not easily scalable and its performance quickly deteroriates when large numbers of clients are added, this is refered to as the Complexity problem in this paper. Lastly DC net protocol has no way of checking the integrity of the clients or the server, thus it has problems with its integrity.*

This paper presents shows how these issues can be resolved, and the research provides an implementation of the DC-net protocol that is practical to deploy and. The application represents a proof of concept for a pure client-server implementation of the DC-net protocol, which avoids the complexity problem found in the implementation scenarios.

VI. FUTURE WORK

Numerous improvements to the implementation can be made with regard to security. To increase the security of the protocol and reduce the chance that a malicious third-party can intercept information with which to compromise the security of the protocol, the primary communication channel used specifically for the DC-net protocol can be encrypted. Currently the server has no ability to police the clients on the server.

Functionality for administration of the server should be added to a production quality implementation to allow the administrators to remove and ban users, among other possible functions. Further improvements to add production quality to the server would be the development of a GUI for the server. A malicious disruption detection method should be implemented to detect users who use customized clients designed for disrupting the DC-net communication on the server. Once identified, these users can be forcibly ejected from the server and their IP address may be banned.

REFERENCES

[1] Nissenbaum, H. 1999. The Meaning of Anonymity in an Information Age. The Information Society , 15:141-144

[2] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. 2007. l-diversity: Privacy beyond K-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) , 1 (1), 3.

[3] Li, N., T. Li, and Venkatasubramanian, S. 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. International Conference on Data Engineering (ICDE) , pp.106-115.

[4] R. Collier, C. Fobel, L. Richards, and G. Grewal A Formal and Empirical Analysis of Recombination for Genetic Algorithm Based Approaches to the FPGA Placement Problem IEEE Canadian Conference on Electrical and Computer Engineering, Montreal, Canada, 2012.

[5] 5. Ohm, P. 2010. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, UCLA LAW REVIEW 57 p. 1701 - 1777. Available from: http://papers.ssrn.com/sol3/papers.cfm ?abstract_id=1450006. Retrieved October 1, 2013.

[6] Sweeney, L. 2002. 'K-anonymity: A Model for Protecting Privacy.' International Journal of Uncertainty, Fuzziness, and Knowledge-based Systems 10(5): 557-570.

[7] Arvind, Narayanan and Vitaly Shmatikov. 2008. Robust De-Anonymization of Large Sparse Datasets, in Proc. of the 2008 IEEE Symp. On Security and Privacy 111 -121.

[8] Akdeniz, Y. 2002. Anonymity, Democracy, and Cyberspace. Social Researc h, 69(11), 223-237.

[9] Sharp J. M., Specialist in Middle Eastern Affairs. Egypt: Background and U.S. Relations. June, 2013. http://www.fas.org/sgp/crs/mideast/RL33003.pdf. Retrieved: October 1, 2013.

[10] Ryan, Yasmine (26 January 2011). "How Tunisia's revolution began – Features". Al Jazeera. Retrieved 13 February 2011. http://www.aljazeera.com/indepth/features/2011/01/2011126121815985483.html. Retrieved: October 1, 2013.

[11] Heydemann S. Syria's Uprising: sectarianism, regionalization, and state the Levant. Fride and Hivos, 2013. http://www.fride.org/download /WP_119_ Syria_Uprising.pdf. Retrieved: October 1, 2013.

[12] Kelly, Brian (2012). "Investing in a Centralized Cybersecurity Infrastructure: Why 'Hacktivism' can and should influence cybersecurity reform". Boston University Law Review 92 (5): 1663–1710. http://www.bu.edu/law/central/jd/organizations/journals/bulr/volume92n4/documents/KELLY.pdf. Retrieved October 1, 2013.

[13] Hancock, Jeffrey T. and Beaver, David I. and Chung, Cindy K. and Frazee, Joey and Pennebaker, James W. and Graesser, Art and Cai, Zhiqiang. Behavioral Sciences of Terrorism and Political Aggression, 2010 vol 2:2 pp 108 - 132. Retrieved October 1, 2013.

[14] Chaum, David. "The Dining Cryptographers Problem: Unconditional Sender and RecipientUntraceability." Journal of Cryptology, 1988: 65-75.

[15] Chaum, David. "Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms. "Communications of the ACM, 1981: 84-88.

[16] S0cketka, (2011, May 30). Dining cryptographers problem [online]. Available: http://en.wikipedia.org/wiki/Dining_cryptographers_problem

[17] Oracle's Java Tutorials, Lesson: All About Sockets [online]. Available: http://download.oracle.com/javase/tutorial/networking/sockets/index.html.

[18] Simple Java Chat Client-Server ,(2008, January 14). Client-server chat download[online].Available: http://breakdesign.blogspot.com/2008/01/simple-java-chat-client-server.html

High Performance Color Image Processing in Multicore CPU using MFC Multithreading

Anandhanarayanan Kamalakannan
Central Electronics Engineering Research Institute
Chennai Centre, CSIR Madras Complex, Taramani
Chennai – 600113, Tamil Nadu, India

Govindaraj Rajamanickam
Central Electronics Engineering Research Institute
Chennai Centre, CSIR Madras Complex, Taramani
Chennai – 600113, Tamil Nadu, India

***Abstract*—Image processing is an engineering field where stored image data is readily available for parallel processing. Basically data processing algorithms developed in sequential approach are not capable of harnessing the computing power of individual cores present in a single-chip multicore processor. To utilize the multicore processor efficiently on windows platform for color image processing applications, a lock-free multithreading approach was developed using Visual C++ with Microsoft Foundation Class (MFC) support. This approach distributes the image data processing task on multicore Central Processing Unit (CPU) without using parallel programming framework like Open Multi-Processing (OpenMP) and reduces the algorithm execution time. In image processing, each pixel is processed using same set of high-level instruction which is time consuming. Therefore to increase the processing speed of the algorithm in a multicore CPU, the entire image data is partitioned into equal blocks and copy of the algorithm is applied on each block using separate worker thread. In this paper, multithreaded color image processing algorithms namely contrast enhancement using fuzzy technique and edge detection were implemented. Both the algorithms were tested on an Intel Core i5 Quad-core processor for ten different images of varying pixel size and their performance results are presented. A maximum of 71% computing performance improvement and speedup of about 3.4 times over sequential approach was obtained for large-size images using four thread model.**

Keywords*—Color image; fuzzy contrast intensification; edge detection; lock-free multithreading; MFC thread; block-data; multicore programming

I. Introduction

Machine vision systems used in various industrial applications are capable of capturing high resolution images and demands time efficient parallel data processing algorithms in real-time environment. To reduce the processing time of the algorithm on these images, parallel computing in multicore architecture is a well known approach [1]. Different parallel programming libraries such as OpenMP and Message Passing Interface (MPI) are widely applied in the development of parallel image processing algorithms. The authors N.E.A. Khalid et al [2] have implemented parallel multicore sobel edge detection algorithm using MPI and observed that parallel processing performs better than sequential processing in terms of execution speed. Chen Lin et al [3] have proposed a parallel method to perform medical image registration using OpenMP and concluded that multithreading approach saves nearly half of the computing time. Alda Kika and Silvana Greca illustrated the development of multithreaded algorithms for contrast, brightness and steganography applications using Java package and tested their performance on different single-core and multicore processors [4]. The authors concluded that the performance of the complex image processing algorithm on multicore CPU can be improved using multithreaded programming.

In our work, we studied the development of multithreaded C++ algorithms for processing low and high resolution color images on a multicore CPU without using parallel programming library and any other additional hardware. To ensure fine grain (data level) parallelism [5] and computation load balance of the algorithm in a multicore CPU, a lock free multithreaded block-data parallel approach is proposed. In this approach, the image data is shared equally among worker threads and each one manipulates its portion of data.

In VC++ programming, MFC library provides powerful threading Application Programming Interfaces (APIs) [6] for developing concurrent or multithreaded windows based software programs. Multithreaded color image processing algorithms namely contrast enhancement using fuzzy technique and edge detection were developed in Intel Pentium dual-core personal computer and tested on Intel Core i5 CPU. The algorithms were applied on ten selected color image samples of varying size and their execution results are presented. The performance results show that both the algorithms in four thread model attained a speedup of about 3.4 times compared with the sequential approach and saves nearly 71% of algorithm execution time.

The paper is arranged as follows: In section II, MFC multithreading and its application in high performance image processing is described. Section III explains the materials, methods and the color image processing techniques followed in this paper. The performance results of the thread model based parallel algorithms are discussed in section IV. The conclusion is given section V.

II. Multithreaded Image Processing using MFC

MFC is a Microsoft's C++ class library for windows programming. It distinguishes two types of threads namely user interface thread and worker thread [7]. The main use of worker thread is to perform background computation work and it is created by defining the task it should perform. This is done by the declaration of thread function according to the MFC definition. The call function AfxBeginThread() launches

the worker thread [8] and it accepts parameters, which includes thread function name, input to the thread, thread priority and few other required parameters.

In block-data parallel processing, image region is identified as several blocks of data. The source image data is partitioned vertically or horizontally into multiple large blocks with equal size [9,10]. In our thread model based parallel approach, each thread exclusively performs image processing task on individual image data block as shown in the concurrency model Fig.1. To maintain load balance within threads, it is good to consider the number of image blocks equal to number of worker threads [11]. Since the image data is stored and accessed through global variables no message passing or explicit data access control is required between threads. This makes thread definition simple without data locking mechanism [12].

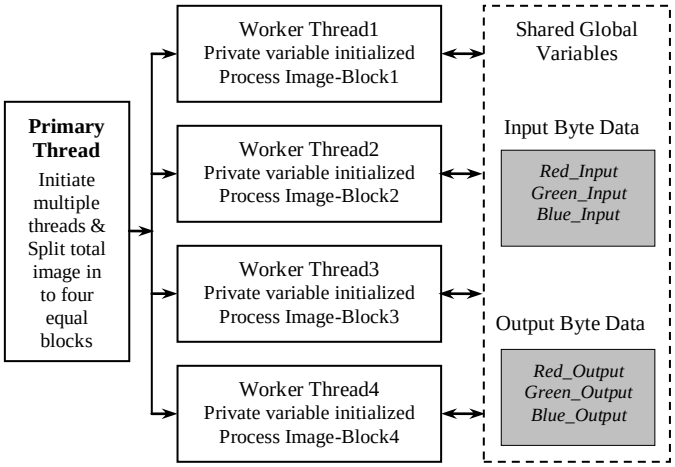


Fig. 1. Concurrency model of block-data parallel algorithm using 4 threads

In this lock free multithreaded approach, threads are free to read and process their portion of image data in a parallel manner, which efficiently reduces the data access time as well as the overall computation time of image processing algorithm. Thus the performance of multithreaded algorithm on a single- chip multicore processor can be fine tuned using shared image data variables [13]. In the case of color image processing algorithm, three input and three output global variables were assigned to each color component (viz. red, green and blue) to enable image reading and processed data writing concurrently using multiple worker threads.

According to the worker thread priority, the operating system schedules each thread to an individual processing unit in a multicore CPU. Due to this scheduling mechanism, all threads do not finish at the same time, so in order to handle this thread completion task, event object is derived from CEvent MFC class. When the thread completes its processing task, the event object is triggered. Using WaitForSingleObject API, event object trigger is noted and worker thread completion is indicated to the primary main thread [6,14,15]. As soon as all the threads complete their processing task, the results are cached and made available in the shared global variable. The synchronization between the main thread and different worker threads was established using event object as shown in Fig.2.

In MFC multithreading, two threads cannot manipulate the same object because MFC objects are thread-safe only at the class level [16]. Hence each thread requires separate objects of the same data structure to operate in a thread-safe manner. To ensure thread safety in the algorithm, each copy of thread parameter data structure is passed as input argument to the corresponding worker thread function. Each thread function uses call by reference method to access the global variables. When the thread calls a image processing function, the private variables declared within the function takes care of storing, processing the intermediate data and also ensures the algorithm execution in parallel manner.

```
//data structure for thread-parameter
typedef struct ThreadParameter
{       int height1, height2, width;
ImageProcessing Object;
} Parameter;

//Global declaration of event object
CEvent threadone, threadtwo;

//Worker thread creation
MainThread()
{
//assign worker thread parameter for image-block1
        Parameter *First=new Parameter;
//assign worker thread parameter for image-block2
        Parameter *Second=new Parameter;
AfxBeginThread(ThreadProcA,First,THREAD_PRIORITY_HIGHEST);
AfxBeginThread(ThreadProcB,Second,THREAD_PRIORITY_HIGHEST);
 ::WaitForSingleObject(threadone,INFINITE);
 ::WaitForSingleObject(threadtwo,INFINITE);
}

//Function Definition_FirstThread
UINT ThreadProcA(LPVOID param)
{       Parameter * ptr=(Parameter *)param;
        ptr->Object.ProcessImage(image-block1);
        threadone.SetEvent();
        return 0;
}

//Function Definition_SecondThread
UINT ThreadProcB(LPVOID waram)
{       Parameter * ptr= (Parameter *)waram;
        ptr->Object.ProcessImage(image-block2);
        threadtwo.SetEvent();
        return 0;
}
```

Fig. 2. Code structure for two thread approach

III. Materials and Methods

A. Sample Images

A total of ten color images with different pixel size were used to evaluate the algorithm performance. All these images were randomly chosen from free online collection of natural scenes and photos. The pixel size of the images varies from 940x474 to 2880x1800. They are labeled as Image1, Image2….Image10.

B. Hardware and Software

The entire coding for developing the multithreaded application software was carried out in Intel Pentium dual-core processor @ 2.8GHz on Windows XP operating system pre-loaded with Microsoft Visual Studio version 6. Using MFC library, multithreaded image processing software for contrast enhancement using fuzzy technique and edge detection has been developed in VC++. Separate menu buttons are provided in the software to load image and execute the algorithms. The developed algorithms were tested using the color image samples on an Intel Core i5-760 @ 2.80 GHz Quad-core CPU on 32 bit Windows 7 operating system with 4GB RAM.

C. Color Image Processing Algorithms

a) Contrast enhancement using fuzzy technique

Image enhancement is a preprocessing technique usually employed to improve the brightness and contrast of the images. In color image enhancement, red, green and blue channels were processed separately and added together to produce composite color value. But this approach does not maintain the color balance in the image. To avoid this change in color information, YIQ color space was chosen, where Y represents the luminance information; I and Q together represent the chrominance information. This color space exploits certain characteristics of human-eye color response and improves the appearance of the color image in terms of human brightness perception. In this technique, the contrast enhancement using fuzzy intensification operator was applied only on luminance component; hence color information of the original image is preserved [17].

Steps involved in image contrast enhancement:

1) Convert RGB image in to YIQ color space [18].

2) Perform fuzzification [19] on luminance component 'Y_{ij}' using the following expression.

$$\mu_{ij} = f(Y_{ij}) = \left[1 + \frac{Y_{max} - Y_{ij}}{f_d}\right]^{-f_e} \quad (1)$$

Where f_e and f_d denote the exponential & the denominational fuzzifier, respectively and μ_{ij} is called the fuzzy property plane of the image. Value of the f_e can be set as 1 or 2. Value of f_d is determined using the cross-over value with respect to fuzziness value 0.5. Y_{max} represents the maximum luminance value.

3) Apply fuzzy intensification operator $T\left(\mu_{ij}'\right)$ on luminance component for contrast enhancement [20].

$$T\left(\mu_{ij}'\right) = \begin{cases} 2\left[\mu_{ij}\right]^2, & 0 \le \mu_{ij} \le 0.5 \\ 1 - 2\left[1 - \mu_{ij}\right]^2, & 0.5 \le \mu_{ij} \le 1 \end{cases} \quad (2)$$

4) Enhanced luminance component 'Y_{ij}' is obtained using defuzzification defined as follows.

$$Y_{ij} = Y_{max} - f_d \left[\frac{1 - \left(\mu_{ij}'\right)^{\frac{1}{f_e}}}{\left(\mu_{ij}'\right)^{\frac{1}{f_e}}}\right] \quad (3)$$

5) Convert YIQ color space to RGB image.

6) Contrast enhanced color image is obtained at the end.

b) Edge detection

Edges in color images can be obtained by applying gray scale edge detection method to each of the RGB bands separately and then results were summed to produce composite value [21]. Further thresholding was performed to get fine binary edges and a set of four Robinson compass masks used in this method are given below.

$$\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{bmatrix} \quad \begin{bmatrix} 2 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & -2 \end{bmatrix}$$

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

D. Multithreaded Block-data Parallel Approach Steps

1) Image block-data decomposition; the main thread splits the image data in to several blocks of equal size to maintain load balance [22].

2) Multiple worker threads are created in the main thread and size parameter of each block is passed as input to the worker thread.

3) Created worker threads are initiated with high priority level to avoid delay due to operating system scheduling.

4) Each worker thread applies its copy of sequential image processing algorithm on a particular image data portion.

5) Each worker thread uses their private copy of data structure for execution.

6) Processed image data are stored in output variable.

7) As shown in Fig.3, main thread exits only when all worker threads complete their assigned task.

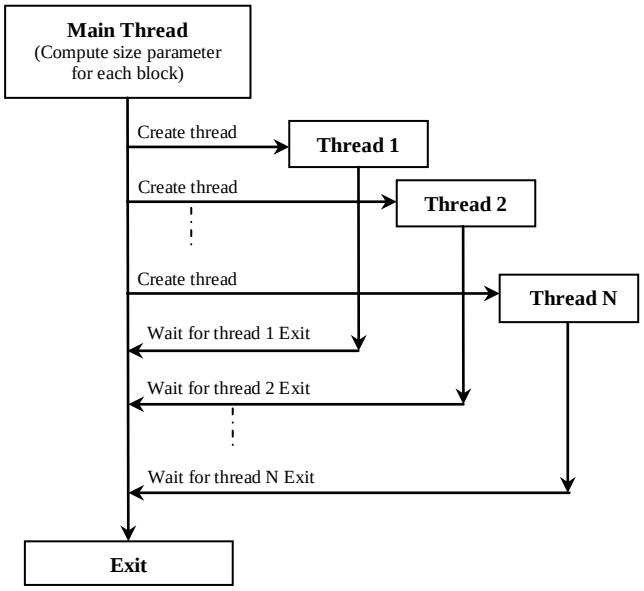


Fig. 3. Thread sequence diagram

IV. Results and Discussion

To measure the execution time of the parallel algorithms while processing the given image data in different threads, the

VC++ clock function was used. Using the execution time, the speedup and performance improvement (P.I) parameters of the algorithms were calculated. The speedup parameter measures how much a parallel algorithm is faster than a corresponding sequential approach [2]. The P.I predicts the relative improvement due to parallel implementation over the sequential approach. The equations for computing the two parameter values are given below.

$$\text{Speedup} = \frac{\text{Sequential\_Approach(ms)}}{\text{Parallel\_Approach(ms)}} \quad (4)$$

$$\text{P.I}\ (\%) = \frac{\text{Sequential\_Approach(ms)} - \text{Parallel\_Approach(ms)}}{\text{Sequential\_Approach(ms)}} \times 100 \quad (5)$$

A. Results of Multithreaded Contrast Enhancement Algorithm

The developed contrast enhancement algorithm was applied on all the ten sample images and the test results were evaluated. The algorithm execution time for each image in sequential and multithreaded approach (for 2, 4 and 8 threads) was recorded in a data file. To determine the average execution time in both the approaches, the algorithm was executed five times successively on each image and the mean time was calculated. The speedup and performance improvement between sequential and four thread approaches was computed using Eq.4 and Eq.5 for images of different pixel size. The algorithm results viz., average execution time, speedup and performance improvement are shown in Table I.

TABLE I. 'Performance Results of Color Contrast Enhancement Algorithm

Image name & pixel size	**Average execution time in milliseconds (ms)**				**Four thread approach**	
	Sequential Approach	*Two Thread*	*Four Thread*	*Eight Thread*	*SpeedUp*	*P.I (%)*
Image1(940 x474)	172	93	68	62	2.53	60.47
Image2(1024 x 728)	265	145	94	93	2.82	64.53
Image3(1280 x 1024)	437	234	140	140	3.12	67.96
Image4(1600 x 1200)	624	328	196	198	3.18	68.59
Image5(1920 x 1080)	675	353	209	212	3.23	69.04
Image6(1920 x 1200)	749	390	231	237	3.24	69.16
Image7(2048 x 1536)	1019	530	312	312	3.27	69.38
Image8(2288 x 1712)	1248	645	375	379	3.33	69.95
Image9(2560 x 1920)	1571	796	458	458	3.43	70.85
Image10(2880 x 1800)	1659	843	483	499	3.43	70.89

To find the maximum possible number of threads needed to speed up the algorithm execution in the Intel Core i5 processor, eight thread approach was also attempted and the execution time results are included in Table I. It is found from the Table I, the execution time of four and eight threads are nearly same which infers that for a quad-core processor, minimum of four MFC thread is enough to achieve optimum execution time.

As seen from the tabulated results, the average execution time of the algorithm decreases with the number of threads, whereas the speedup and performance improvement goes up with increase in image size. In the four thread implementation, the speedup parameter varies from 2.53 to 3.43 times and the performance improvement variation is found to be between 60.47% and 70.89%.

The input image and processed color image outputs of contrast enhancement algorithm are shown in Fig.4a, Fig.4b & Fig.4c. The processed results of sequential and multithreaded approach are looking similar.

Fig. 4. a. Input photograph image

Fig. 4. b. Contrast enhancement in sequential approach

Fig. 4. c. Contrast enhancement in four threads

Table II illustrates the execution time of individual threads measured for contrast enhancement algorithm executed five times successively on a single image. It shows that each thread takes nearly same amount of CPU time to compute the data processing task and the load balance within multiple threads is well maintained. Therefore change in execution time is mainly dependent on varying image size.

TABLE II. Individual Thread Execution Time for Four Thread Contrast Enhancement Algorithm (Image Size: 1920x1200)

Running Iteration No.	**Execution time in milliseconds (ms)**				
	Thread1	*Thread2*	*Thread3*	*Thread4*	*Algorithm*
1	218	234	234	234	234
2	218	218	234	234	234
3	219	234	234	234	234
4	218	218	218	234	234
5	218	218	218	218	218

B. Results of Multithreaded Edge Detection Algorithm

A similar approach as followed for the contrast enhancement algorithm was applied for the edge detection algorithm on all the ten color images and the results are presented in Table III. In four thread approach, the speedup varies from 2.82 to 3.44 times and the performance improvement achieved is between 64.50% and 70.94%.

TABLE III. Performance Results of Color Edge Detection Algorithm

Image name & pixel size	Average execution time in milliseconds (ms)				Four thread approach	
	Sequential Approach	*Two Thread*	*Four Thread*	*Eight Thread*	*SpeedUp*	*P.I(%)*
Image1(940 x474)	307	167	109	104	2.82	64.50
Image2(1024 x 728)	520	265	171	172	3.04	67.12
Image3(1280 x 1024)	837	431	265	255	3.16	68.34
Image4(1600 x 1200)	1229	634	369	369	3.33	69.98
Image5(1920 x 1080)	1304	676	390	405	3.34	70.09
Image6(1920 x 1200)	1466	765	437	442	3.35	70.19
Image7(2048 x 1536)	1992	1024	588	588	3.39	70.48
Image8(2288 x 1712)	2475	1269	728	733	3.40	70.59
Image9(2560 x 1920)	3100	1571	905	899	3.43	70.81
Image10(2880 x 1800)	3276	1648	952	952	3.44	70.94

The input image and processed color image outputs of edge detection algorithm are shown in Fig.5a, Fig.5b & Fig.5c. The processed image outputs of the algorithm are found to be similar.

Fig. 5. a. Input MatLab demo image

Fig. 5. b. Edge detection in sequential approach

Fig. 5. c. Edge detection in four threads

Thus the two multithreaded color image processing algorithms with different complexity levels were tested in Intel Core i5 processor and found that four thread approach utilized the quad-core CPU efficiently on Windows 7 platform.

V. CONCLUSION

This work was carried out to explore the parallel processing ability of the multicore CPU in processing high resolution images using MFC multithreading. In this paper, a lock-free multithreaded block-data parallel approach based color image processing algorithms for fuzzy contrast enhancement and edge detection were developed using VC++ on windows platform without using any parallel programming library. The purpose of this implementation is to improve the performance and reduce the execution time of the image processing algorithms on multicore processor by partitioning the given image into equal blocks and processing each block of data in a parallel manner. In four thread approach, the algorithm speed is found to be about 3.4 times faster than the sequential approach. With regard to performance improvement, the thread model saves nearly 71% computation time compared to sequential implementation. No performance improvement and speedup is noted in processing nearly same size images of marginal difference in pixel size. The performance results indicate that multithreaded image processing algorithms efficiently utilize the computing capability of multicore CPU like Intel Corei5 processor. Hence the developed multicore programming approach using MFC thread can be applied to improve the performance of various color image processing algorithms.

REFERENCES

[1] P.N.Happ, R.S.Ferreia, C.Bentes, G.A.O.P.Costa and R.Q.Feitosa, "Multiresolution Segmentation: A parallel approach for high resolution image segmentation in multicore architectures", International Conference on Geographic Object-Based Image Analysis, ISPRS Vol.XXXVIII-4/C7, June-July 2010.

[2] N.E.A.Khalid, S.A.Ahmad, N.M.Noor, A.F.A.Fadzil and M.N.Taib, "Parallel approach of sobel edge detector on multicore platform", International Journal of Computers and Communications, Vol.5 Issue.4, 2011, pp.236-244.

[3] Chen Lin, Li Jian, Zhou Jun and Jiang Murong, "Multithreading method to perform the parallel image registration", IEEE Xplore, International Conference on Computational Intelligence and Software Engineering, DOI:10.1109/CISE.2009.5366052, Dec. 2009.

[4] Alda Kika and Silvana Greca, "Multithreading image processing in single-core and multi-core CPU using Java", International Journal of Advanced Computer Science and Applications, Vol.4, No.9, 2013, pp.165-169.

[5] Luis Moura E Silva and Rajkumar Buyya, "Chapter1: Parallel programming models and paradigms", High Performance Cluster Computing: Programming and Applications, Vol.2, Prentice Hall PTR, 1999, pp.4-28.

[6] S.Akhter and J.Roberts, "Chapter 5: Threading APIs", Multicore Programming: Increasing performance through Software Multi-threading, Intel Press, 2006, pp.75-133.

[7] J. Prosise, "Chapter 17: Threads and thread synchronization", In Programming Windows with MFC, 2nd Edition, Microsoft Press, 1999, pp.985-1000.

[8] Stanford Taylor Jones and Chi Ngoc Thai, "Multithreaded Design of Spectral Imaging Software", ASAE Meeting Presentation, Paper No.053010, July 2005.

[9] Winser E.Alexander, Douglas S.Reeves and Clay S.Gloster, "Parallel image processing with the block data parallel architecture", IEEE Xplore, Proceedings of the IEEE, DOI:10.1109/5.503297, Vol.84, No.7, July 1996, pp.947-968.

[10] A.Fakhri A.Nasir, M.Nordin A.Rahman and A.Rasid Mamat, "A study of image processing in agriculture application under high performance computing environment", International Journal of Computer Science and Telecommunications, Volume 3, Issue 8, 2012, pp.16-24.

[11] Sanjay Saxena, Neeraj Sharma and Shiru Sharma, " Image processing tasks using parallel computing in multicore architecture and its applications in medical imaging", International Journal of Advanced Research in Computer and Communication Engineering, Volume 2, Issue 4, 2013, pp.1896-1900.

[12] Jonathan R.Engdahl and Dukki Chung, "Lock-free data structure for multi-core processors", International Conference on Control, Automation and Systems, October 2010, pp.984-989.

[13] Devrim Akgun, "Performance evaluations for parallel image filter on multi-core computer using Java threads", International Journal of Computer Applications, Vol.74, No.11, July 2013, pp.13-19.

[14] S.S.Ilic, A.C.Zoric, P. Spalevic and Lj. Lazic, "Multithreaded application for real-time visualization of ECG signal waveforms and their spectrums", Intl. Journal of Computer, Communication & Control, 8(4), 2013, pp.548-559.

[15] Faran Mahmood, "Parallel Implementation of Imaging Filters on Multi-Core Processors for Win32 platform", Proceedings of the 4th International Conference on Open-Source Systems and Technologies, December 2010.

[16] Multithreading: Programming Tips – "Accessing objects from multiple threads", Available from http://msdn.microsoft. com/en-us/library/h14y172e.aspx.

[17] Zhuqing Jiao and Baoguo Xu, "An image enhancement approach using Retinex and YIQ", IEEE Xplore, Inter-national Conference on Information Technology and Computer Science, DOI:10.1109/ITCS.2009.104, July 2009, pp.476-479.

[18] Gwanggil Jeon, "Image enhancement in YIQ space", Proceeding of First International Conference on Advanced Computer and Information Technology, ASTL vol.22, 2013, pp.109-112.

[19] Sankar K.Pal and Robert A.King, "Image enhancement using smoothing with fuzzy sets", IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-11, No.7, 1981, pp.494-501.

[20] Peng Dong-liang and Xue An-ke, "Degraded image enhancement with applications in robot vision", IEEE Xplore, International Conference on Systems, Man and Cybernetics, DOI:10.1109/ICSMC.2005.1571414, Vol.2, October 2005, pp.1837-1842.

[21] Scott E.Umbaugh, "Chapter 4: Segmentation and Edge/Line Detection", Computer Imaging- Digital image analysis and processing, CRC Press, 2005, pp.184-188.

[22] Young-Jip Kim and Byung-Kook Kim, "Load balancing algorithm of parallel vision processing system for real –time navigation", IEEE Xplore, International Conference on Intelligent Robots and Systems, DOI:10.1109/IROS.2000. 895242, Vol.3, Oct-Nov 2000, pp.1860-1865.

FIR Filter Design Using The Signed-Digit Number System and Carry Save Adders – A Comparison

Hesham Altwaijry
Computer Engineering Department,
King Saud University
PO Box 51178, Riyadh 11543, Saudi Arabia

Yasser Mohammad Seddiq
Computer Engineering Department,
King Saud University
PO Box 51178, Riyadh 11543, Saudi Arabia

***Abstract*— This work looks at optimizing finite impulse response (FIR) filters from an arithmetic perspective. Since the main two arithmetic operations in the convolution equations are addition and multiplication, they are the targets of the optimization. Therefore, considering carry-propagate-free addition techniques should enhance the addition operation of the filter. The signed-digit number system is utilized to speedup addition in the filter. An alternative carry propagate free fast adder, carry-save adder, is also used here to compare its performance to the signed-digit adder. For multiplication, Booth encoding is used to reduce the number of partial products. The two filters are modeled in VHDL, synthesized and place-and-routed. The filters are deployed on a development board to filter digital images. The resultant hardware is analyzed for speed and logic utilization**

Keywords*— *FIR Filters – Signed Digit – Carry-Save – FPGA

I. INTRODUCTION

Digital signal processing (DSP) systems employ computer systems to digitally process input signals. An example where computer arithmetic is a key factor in optimizing the design is digital filters especially convolution-based ones. The hardware complexity and processing delay of these digital filters are proportional to a parameter called the filter order, which is highly desired to be large [1]. These filters have been built using FPGA's [2] [3] [4]

A fundamental principle in computer arithmetic upon which all the succeeding aspects are based is how values are to be represented. As all the computing platforms that are used today for digital signal processing are based on digital electronics, the arithmetic operations they perform should be handled in a way that is suitable to the nature of the electronics that build these platforms. The way a value is represented is called a number system. Computers were initially developed to use the binary number system (radix-2). Although, computers use radix-2, there have been few number systems discussed in the computer arithmetic literature that are unconventional in terms of representation and operations. Such number systems are used in computers for some special applications.

A. The Binary Number System

Binary number systems are called positional number systems [5]. A general expression for the value of an n-digit number A consisting of digits $a_{n-1}, a_{n-2}, \ldots, a_0$, in radix-$r$ number system is as follows:

$$A = \sum_{i=0}^{n-1} a_i \times r^i \tag{1}$$

In computers, the choice of r is 2 due to electronic circuit limitations. When r equals a constant value as in the decimal and the binary systems, this is called a fixed-radix number system. An observation on the conventional fixed-radix positional representation is that special representations are required for signed number and that carry propagation in addition, which increases the delay of operations, limits system scalability and adds more complexity to algorithm implementation.

B. Unconventional Number Systems

A common feature of the unconventional number systems is redundancy; a positional number system is redundant when the number of elements in its digit set is greater than r, where r is the radix. In a redundant number system, an algebraic value can have more than one representation. Redundant number systems can improve system reliability, increase speed of operations, and provide structural flexibility**.** [6]

Signed digit number systems are a positional number representation with a constant radix $r \geq 3$. Each digit of a signed-digit number can have one value of the set $\{-a, -a+1, \ldots, 0, \ldots, a-1, a\}$ [7] [8]. The maximum possible magnitude, a, is set as follows

$$\frac{r_o+1}{2} \leq a \leq r_o - 1 \qquad \text{for odd radices } r_o \geq 3 \tag{2}$$

$$\frac{r_e}{2} + 1 \leq a \leq r_e - 1 \qquad \text{for odd radices } r_e \geq 4 \tag{3}$$

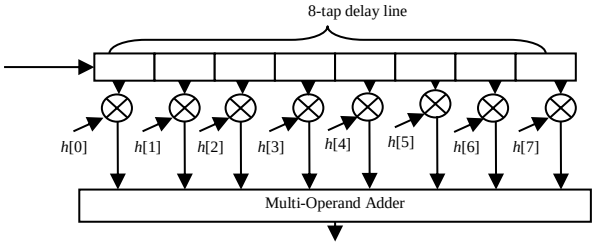


Fig. 1. Basic FIR Filter

When a value is represented with *n* binary bits, then it will be represented with $k = \lceil n / \lceil \log_2 r \rceil \rceil$ signed digits [9]

Signed-digit systems have the advantage that the addition time of a multi-operand adder that is built by cascading identical digit adders is constant.

A special case in signed-digit representation is for the radix *r* = 2 and the digit set is { $\overline{1}$, 0, 1} where $\overline{1}$ represents – 1. In this case, the representation is called canonic signed-digit (CSD). Three main properties of the CSD are that it is irredundant, the number of nonzero digits is minimal and multiplying any two adjacent digits will produce zero. In the applications that involve multiple constant multiplications (MCM) as in the FIR filters, using CSD guarantees the minimal number of adders.

II. Digital Filters

Filters are signal processing components that are used to process interfered and corrupted signals. They can be classified to two main categories: analog and digital filters. Filters in these two categories are different in terms of cost, speed, accuracy, power consumption and implementation, but they are similar in the sense that they are both used to filter signals.

A commonly used method of implementing digital filters is by considering a subset of the filter's impulse response. Filter designed this way are called finite impulse response (FIR) filters. The mathematical process used to get the output of a linear system according to its impulse response is the convolution. When a digital signal *x*[*n*] is to be processed by a system of impulse response *h*[*n*], the output is the result of the following equation [1]:

$$y[n] = \sum_{k=0}^{N-1} h[k]x[n-k] \tag{4}$$

The above equation describes how each sample of the output signal is calculated. This is an application of the widely used mathematical operation of the dot product, which consists purely of multiplication and addition. Optimizing the dot product does not only serve the FIR filter application, but also some other applications that are similarly described such as radar processing, signal correlation and matrix multiplication.

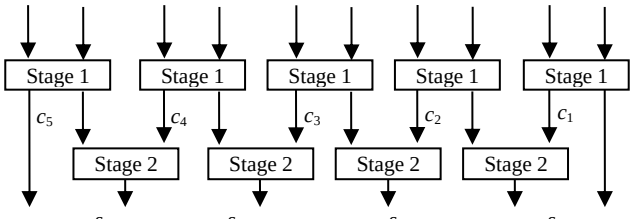


Fig. 2. Signed Digit Addition

A general block diagram of the convolution process as it is implemented in hardware is shown in Fig. 1 The delay line represents the inversion and shift in the input *x*[*n*]. The taps of the delay line are multiplied by the constant values of *h*[*n*]. The wider the delay line, the more accurate the results of the FIR filter are. Of course, this is on expense of more hardware resources, higher power consumption and higher cost.

In order to make the FIR filter performs addition faster, breaking the carry propagation chain in its adders is essential. The two most common techniques to achieve this are signed-digit addition and carry-save addition.

A. Signed-Digit Addition

Signed-digit number system can perform addition and subtraction with a limited propagation of carry. This feature makes the adder's delay independent of the operand length which implies less delay. The carry propagation can go as far as one position to the left. The sign of the number is implicitly expressed in the digits and no special representation is needed for this purpose. A block diagram of a signed digit adder is depicted in Fig. 2. The values shown in the figure are determined in [10]

B. Carry-Save Addition

Carry-save addition is one of the carry-propagate free methods of addition [11]. Carry-save adders (CSA) are mainly used when adding three operands or more. CSAs are built using (3, 2) counters in a manner that prevents carry propagation. The term (3, 2) counter is an alternative name for a full adder because it receives three bits of the same weight and outputs two bits representing the number of ones in the three-bit input. In the ordinary carry-propagate adders, the least significant bit of the output of the (3, 2) counter is the sum while the most significant bit is the carry that propagates to the left. Therefore, adding multi-operands using a CSA will result in a vector of the sum bits and another vector of carry bits. This two-vector result invites the need for a carry-propagate adder to add these two vectors to get a single result in the normal binary representation.

III. Implementing Signed Digit Filter

For the purpose of implementing a high-speed FIR filter, the arithmetic advantages of the signed-digit number system have been exploited to enhance the filter performance. This section is an elaboration to our work in [12]. It discusses the adder and the multiplier design and implementation.

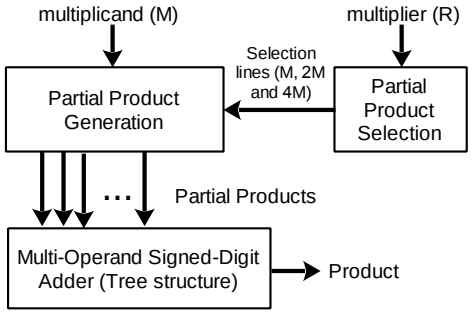


Fig. 3. Signed Digit Booth-3 Multiplier

A. Digit Set and Encoding

The signed-digit number system used in this work is in radix-2. Therefore, the allowed set of digits is $\{\bar{1}, 0, 1\}$ where $\bar{1}$ represents –1. As there are three possible digits in this set, two bits are needed to encode each digit.

There are two commonly used encoding methods used to encode the digits of a number in the signed-digit representation [10] [13]. In the first method, each digit is assigned a 2's complement number representing the algebraic value of that digit. In the second encoding method, each digit of the number in signed-digit representation is assigned a 2-bit code x such that the sum of the two bits is equal to the value of that digit. If the high bit holds a negative sign, the low bit holds a positive sign and vice versa. That is, the value can be determined either as $x^+ + x^-$ or as $x^- + x^+$.

The second encoding method makes converting signed-digit numbers to 2's complement number easier [10]. Additionally, sign inversion of a digit is simply swapping the high and the low bits. Therefore, in this work the second encoding method with the value determined as $x_h - x_l$ is used.

A Signed digit adder can be implemented as a straight forward implementation into an FPGA by means of using lookup tables (LUT). However, the synthesis of this implementation results in a very poor utilization for the FPGA logic elements. Alternatively, the adder can be implemented using logic gates based upon equations presented by [14].

B. Signed Digit Partial Product Generation

FIR filters involve multiplying the input samples with the filter kernel coefficients. Thus, improving the filter multipliers will significantly improve the filter performance. Multiplication is done two steps:

1. Generating the partial products.
2. Accumulating the partial products.

The number of partial products needed for the multiplication can be reduced by using Booth's algorithm [15], however this encoding is performed serially, it can be done in parallel using the modified Booth encoding [16]. Booth-2 encoding is the most commonly used method. However, Booth-3 provides more reduction for the non-zero digits but the existence of the hard multiple 3M forms an obstacle when applying Booth-3 encoding. An efficient solution for the hard multiple 3M was proposed in [13] by exploiting the advantages of the signed-digit number system. The multiplier proposed accepts two operands in 2's complement representation and gives their product in signed-digit representation. The digit coding method used in that work is the sum-of-bits in the form $x^+ + x^-$. The signed-digit encoding method is utilized to determine the hard multiple 3M as 4M – M.

Finally, the partial products are added up to get the final product. This step requires a signed-digit multi-operand adder. Binary tree architecture is used to build the multi-operand adder using two-operand signed-digit adders. The block diagram of the Booth-3 multiplier that is designed in this work is shown in Fig. 3

C. Signed Digit Filter

After the multiplier and the multi-operand adder are already implemented, the FIR filter just needs to be assembled. The delay line has been implemented as an array of registers. Since the output of the final adder is still in signed-digit representation, a converter had to be added to convert the result to 2's complement format. In this work, the converter is simply a carry lookahead adder that adds the positive and the negative parts of the signed-digit number. Converting an n-digit signed-digit number X is performed as follows:

$X = (x^+, x^-)_{n-1} (x^+, x^-)_{n-2} \ldots (x^+, x^-)_2 (x^+, x^-)_1 (x^+, x^-)_0$.
The positive part is $XP = x^+_{n-1} x^+_{n-2} \ldots x^+_2 x^+_1 x^+_0$.
The negative part is $XN = x^-_{n-1} x^-_{n-2} \ldots x^-_2 x^-_1 x^-_0$
The 2's complement representation is $Y = XP + XN$.

A scalable and parameterized design has been highly considered. Thus, when the FIR filter is assembled, the filter order and the sample width are defined as design generics in the VHDL code such that the generated filter architecture meets the intended filter parameter. The block diagram of the signed-digit FIR filter implemented in this work is depicted in Fig. 4. If the sample width is *W* and the filter order is *N*, the filter output sample will be of width $2W + \lceil \log_2 N \rceil$.

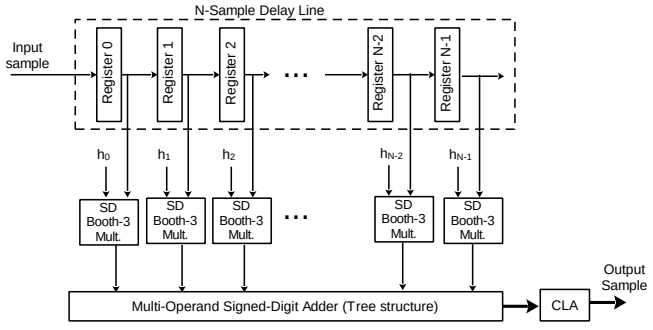


Fig. 4. Signed Digit FIR Filter [12]

IV. Implementing Carry Save Filter

In this section, another method of breaking the carry propagation chain is reported. That is by using carry-save addition (CSA) [11]. CSAs are built using (3, 2) counters in a manner that prevents carry propagation. The term (3, 2) counter is an alternative name for the full adder because it

receives three bits of the same weight and outputs two bits representing the number of ones in the three-bit input. In CSA instead of propagating the carry bits to a higher position, these carry bits are kept and added using later stages of the CSA. Carry-save adders are used in FIR filters to add the partial products of the multipliers and to calculate the final result of the filter.

A. Carry-Save Addition

An efficient way of designing a carry-save adder to achieve fast performance is by designing it based on a 3-operand carry-save adder. A k-operand CSA adder (where k > 3), is constructed out of several blocks of 3-operand CSAs [17] [18]. This k-operand CSA could be implemented in two common ways: cascade or tree. The cascade structure accepts one new operand at each level except at the first level where three new operands are accepted. The number of levels in this structure is more than the number of levels in the tree structure, which implies more delay. However, the cascade structure remains a preferred option sometimes due to its regular layout, which implies more simplicity in the VLSI design.

On the other hand, the tree structure, which is known as the Wallace tree [11], accepts as many operands as possible at the first level. The following levels are used to add the sum and the carry vectors in addition to the operands remaining from the first level, which must be at most two remaining operas. When using the tree structure to build a CSA, the number of levels will be less than the cascade structure. [19]

B. Carry-Save Filter

The CSA and Booth-2 multiplier are the fundamental blocks of the carry save FIR filter. As in the signed-digit case, the FIR filter is built by assembling these blocks with some extra logic for the delay line, which is an array of registers, and the converter, which is a carry lookahead adder. The conversion from the carry-save to the 2's complement format is performed by adding the sum and carry vectors. The filter is illustrated in Fig. 5

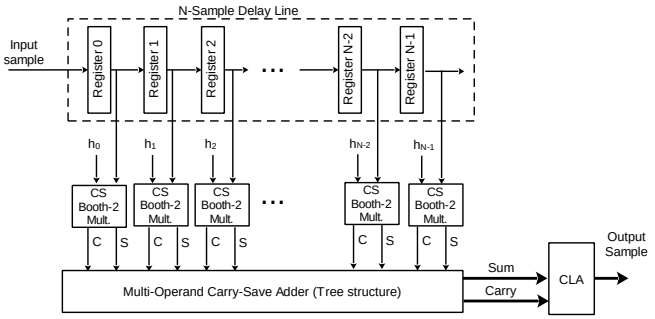


Fig. 5. Carry-Save FIR Filter

V. Testing and Verification

The implementation of the different configurations of the FIR filters have been tested functionally using test benches written in VHDL. Those test benches covered the top-level architectures along with the sub-components throughout the design hierarchy. Moreover, the filters have been synthesized and mapped into an FPGA in order to verify their functionality on real hardware for image processing. This experiment is reported briefly in [12] while described in more details in the following. Such applications are challenging in the sense that two dimensional (2D) FIR filters are needed instead of the one dimensional (1D) filters available in hand. A 2D FIR filter applies the 2D convolution equation:

$$y[m,n] = \sum_{i=0}^{M-1}\sum_{j=0}^{N-1} x[i,j]h[m-i,n-j] \tag{5}$$

A sample of an image signal is in indexed by two values: *m* and *n* indicating the row and column position of that sample respectively. Since the filters implemented so far in this work are 1D, they are instantiated in parallel to build a 2D FIR filter of order M-by-N such that each one of the M 1D filters performs convolution operation of order N as illustrated in Fig.6. The kernel h of each filter is assigned as one row of the 2D kernel. The 2D filter of order 11-by-11 that is designed in this work is intended to smooth out the sharp edges of an input image by averaging out each pixel with its neighbors such that the output image is a blurred version of the original one. So, this is a moving average filter, which is a low-pass filter. After blurring the image, the result is subtracted from the original image in order to extract the image edges and cancel out the constant regions. In this work the filter size is 11-by-11 and the filter kernel is

$$h = \begin{bmatrix} \frac{1}{11^2} & \frac{1}{11^2} & \frac{1}{11^2} & \cdots & \frac{1}{11^2} \\ \frac{1}{11^2} & \frac{1}{11^2} & \frac{1}{11^2} & \cdots & \frac{1}{11^2} \\ \frac{1}{11^2} & \frac{1}{11^2} & \frac{1}{11^2} & \cdots & \frac{1}{11^2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{11^2} & \frac{1}{11^2} & \frac{1}{11^2} & \cdots & \frac{1}{11^2} \end{bmatrix} \tag{6}$$

The above described process is implemented on Altera Cyclone II Starter Development Board. The image source is a normal personal computer. The interface between the FPGA and the computer is the serial port. The results are depicted in Fig. 7

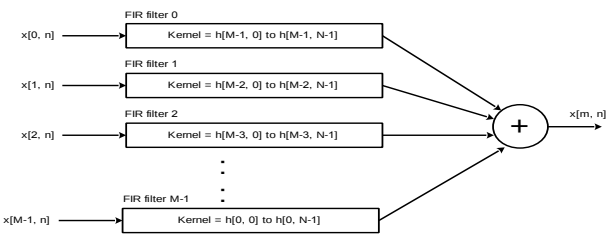
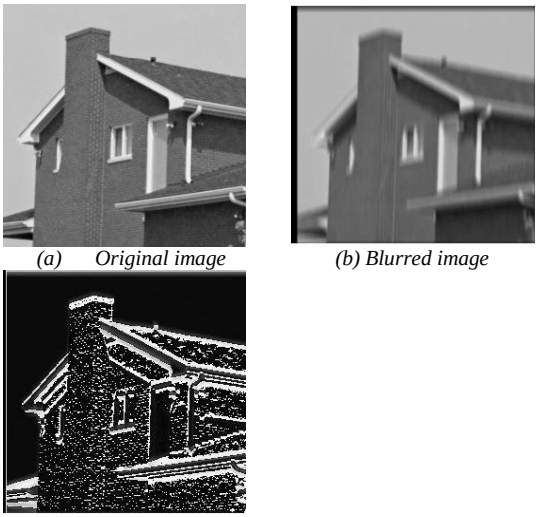


Fig. 6. 2-D FIR filter [12]



(a) Original image *(b) Blurred image*

(c) Subtracted image

Fig. 7. Edge Extraction using FIR filter [12]

VI. SYNTHESIS RESULTS

The two FIR filter designs have been implemented in a scalable and parameterized manner by exploiting the generality features of VHDL. The sample width W and the filter order N are the two generics of the two filters. Recalling that the major problem that is handled in this work is the carry propagation delay, different values for W, which is directly affecting the carry chain, should be examined. Since we are talking about FIR filters, the order N is also worth examining.

Three precision levels of W are selected: low precision, medium precision and high precision where W is equal to 12, 24 and 48 bits respectively. These values of W are selected because when Booth-3 is applied on them, the number of generated partial products is a power of 2, which reduces the complexity of the multi-operand adder when built as a binary tree. In fact, there is also some attractive and practical advantage for choosing sample widths of 12 and 24 bits. That is, most of the commercial analog-to-digital converters (ADCs) that are used today electronic systems are 12-bit wide and most of the audio codec components used in media systems are 24-bit wide.

For the values of N, arbitrarily chosen values of 16, 32 and 64 samples have been considered. With two filter types, three precision levels and three filter orders, there are 18 different FIR filters to be synthesized. The 18 filters have been synthesized and place-and-routed using Quartus II software. The target FPGA is Altera Stratix III EP3SL340 [20]. The timing and hardware results of the place-and-route process are targeted for analysis. To get accurate timing results, the delay from the primary input to the primary output of the filter should be measured. In fact, some software design tools measure the delay starting from the FPGA pin to which the primary input is assigned and ending at the pin to which the primary output pin is assigned. This method of measuring delay is not accurate when comparing two or more designs because there will be some extra routing delay between the pins of the primary ports. This routing delay is dependent on where the tool places and routes the design and hence is not regular. To avoid this problem in this work, the primary input and the primary output of the filter are latched. Once the tool figures out that there is some logic between two registers, it will calculate the maximum clock frequency fmax allowed for this design. The reciprocal of the frequency (1/ fmax) is the propagation delay of the logic between the two registers, plus some sequencing overhead which is common for all designs. Thus, the presence of this extra delay in the comparison is fair. The place-and-route results for delay and hardware are collected and analyzed. These results are discussed in the following figures.

The delay data collected for the 18 filters are summarized in Fig.8. An observation on the chart is that the latency in both filters is proportional to W and N. The justification of this observation is that the sample width W directly affects the number of partial products that is generated in the Booth multipliers, which in turn increases the number of levels in the multi-operand adder inside the multiplier. Likewise, the value of N affects the number of levels in the multi-operand adder that generates the final filter result. This increase in the adder levels, which is proportional to W and N, results in a larger latency. It is interesting to notice that the delay of the two filters is not equally proportional to W and N; it is highly proportional to W while slightly proportional to N. This is, in fact, due to the delay introduced by the carry lookahead adder that acts as a converter at the last stage in both filters. A CLA is a carry propagate adder and it is expected to be highly tied to W. When this CLA has been separately analyzed, it has been found that it is responsible for about 28 % of the overall filter delay. This explains why W has more influence on the filter delay than N.

Another observation is that the signed-digit filter is always faster than the carry-save filter. This improvement in filter performance needs more analysis in order to see how the improvement behaves with respect to the design parameters and how significant it is. Fig. 9 depicts the ratio of the signed-digit filter delay over the carry-save filter delay. Clearly, the delay improvement is very slight since the chart indicates that signed-digit filter performs between 1.1 and 1.2 times faster than the carry-save filter. The speedup is almost constant regardless of the values of W and N.

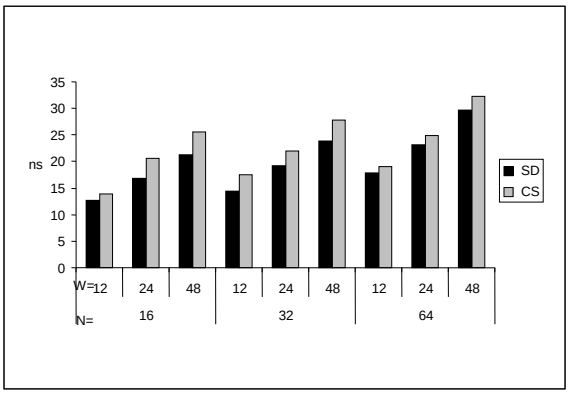


Fig. 8. FIR filter Delays

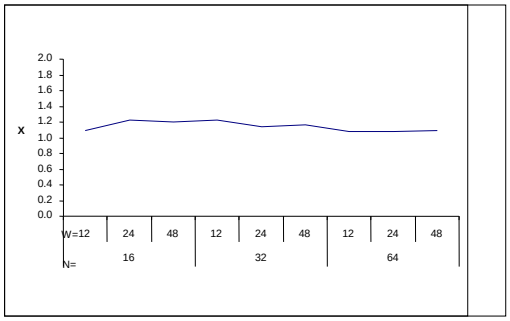


Fig. 9. Ratio of SD- to CSA FIR filter delays

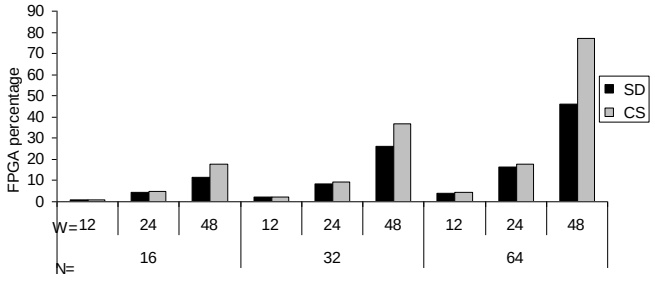


Fig. 10. Logic Utilization

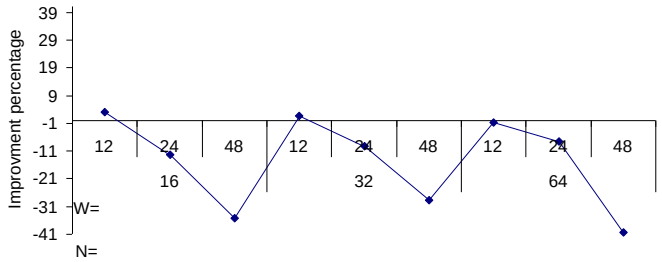


Fig. 11. Logic Utilization Percentage

The data of the logic utilized by the 18 filters is also collected and analyzed. The place-and-routing logic utilization results are summarized in Fig. 10. While Fig. 11 shows the percentage of hardware reduction in the signed-digit filter with respect to the carry-save filter. It is logical and expected to see that the hardware of the two filters grows as the values of W and N increase. It is noticeable that the hardware utilization when W = 12 bits, regardless of the value of N, is almost the same for the two types of filters. With higher values of W, the signed-digit filter has an advantage. The case of having the signed-digit filter being smaller than the carry-save filter despite that the added complexity of the signed-digit adder is may by the significance of Booth multipliers in the filters since most of the filter size is occupied by them. Booth multiplier efficiency in saving hardware and time becomes more significant and appreciated when the multiplier gets bigger. This is what makes the logic difference more notable when W is equal to 24 and 48 bits. From Fig. 11, the signed-digit filter is between 30 % and 40 % smaller than the carry-save filter for W = 48. This might seem counterintuitive. However, this could be justified by amount of multipliers used, and the fact that the reduction from Booth 2 to Booth 3 is the cause for this reduction in size.

VII. Conclusions

In this work, FIR filter design and implementation have been approached from arithmetic perspective. The signed-digit and the carry-save arithmetic techniques have been exploited to reduce addition time. Booth encoding was used to speedup multiplication. The authors designed, simulated and tested a high-speed FIR filter using the signed digit number system. The first part of work is the design and implementation of the FIR filter using the signed-digit number system and Booth-3 encoding to improve the filter adders and multipliers respectively. The implementation of the signed-digit two-operand and multi-operand adders has been discussed. In Booth-3 implementation, it has been shown how the signed-digit number system helps in generating the hard multiple 3M. The other part of this work is the design and implementation of an FIR filter using carry-save addition and Booth-2 encoding to improve the filter adders and multipliers respectively. The hierarchical design of CSAs of several sizes has been reported.

For the two types of filters in this work, nine different configurations have been considered for the sample width W (12, 24, 48 bits) and for filter order N (16, 32, 64) samples. A total of 18 filters of both types have been modeled and generated in VHDL. Then, these filters have been synthesized and place-and-routed. The data resulted from the place-and-rout process, which is related to system delay and logic size, has been collected and analyzed.

The results analysis have shown that the signed-digit FIR filters designed in this work are slightly faster than the carry-save FIR filters. The filter delay is slightly proportional to N, but highly proportional to W. The signed-digit filters are constantly about 1.1 times faster than the carry-save filters. Both types of filters have consumed almost the same amount of logic for low precision samples while they differ in logic utilization as the precision increases. The signed-digit filter reported better logic utilization especially for W = 48 bits where it becomes 30 % to 40 % smaller than the carry-save filter.

In conclusion, both the signed-digit and the carry-save filters are fast and efficient because of the carry-propagate-free addition they involve. The speedup that is gained in the FIR filter when signed-digit arithmetic is used is not so significant. Likewise, the filter size reduction for a sample width around 12 bits is almost negligible. However, the improvement in logic utilization for wider samples is strongly significant. Therefore, designing FIR filters using signed-digit number system becomes efficient and useful more than carry-save filters when the filter works for high precision samples. However, the signed-digit filter is superior over the carry-save filter in logic utilization more than speed.

References

[1] S. Smith, The Scientist and Engineer's Guide to Digital Signal Processing, San Diego: California Technical Publishing, 1997.

[2] C.-J. Chou, S. Mohanakrishnan and J. B. Eva, "FPGA Implementation of Digital Filters," in International Conference of Signal Processing Applications and Technology ICSPAT '93, Santa Clara, CA, 1993.

[3] B. Parhami and D.-M. Kwai, "Parallel Architectures and Adaptation Algorithms for Programmable FIR Digital Filters with Fully Pipelined Data and Control Flows," Journal of Information Science and Engineering, no. 19, pp. 59-74, 2003.

[4] X. Jiang and Y. Bao, "FIR filter design based on FPGA," in International Conference on Computer Application and System Modeling (ICCASM) , Taiyuan, 2010.

[5] J. Deschamps and M. Davio, "Addition in Signed Digit Number System," in Proceedings of the eighth international symposium on Multiple-valued logic , Rosemont, Illinois, United States , 1978.

[6] D. Atkins, "Introduction to the Role of Redundancy in Computer Arithmetic," Computer, vol. 8, no. 6, pp. 74-77, June 1975.

[7] A. Avizieni, "Binary Compatible Signed Digit Arithmetic," AFIPS Conference Proceedings, vol. 26, no. 1, pp. 664-672, 1964.

[8] P. Ramamoorthy, B. Potu and G. Govind, "DSP System Architecture Using Signed-Digit Number Representations," ICASSP, vol. 3, pp. 1702-1705, April 1988.

[9] C. Nagendra, M. Irwin and R. M. Owens, "Area Time Power Tradeoffs in Parallel Adders," IEEE Transaction on Circuits and Systems, vol. 43, no. 10, pp. 689-702, October 1996.

[10] I. Koren, Computer Arithmetic Algorithms, Natick: A. K. Peters, 2002.

[11] S. Wallace, "A Suggestion for a Fast Multiplier," IEEE Transactions of Electronic Computers, pp. 14-17, Febuary 1964.

[12] Y. M. Seddiq and H. A. Altwaijry, "An Implementation of a 2D FIR Filter Using the Signed-Digit Number System," in *Saudi International Electronics, Communications and Photonics Conference (SIECPC2011)*, Riyadh, pp. 1-4. 2011.

[13] O. McSorley, "High Speed Arithmetic in Binary Computers," Proceedings of the IRE, vol. 49, no. 1, pp. 67-91, January 1961.

[14] H. Makino, Y. Nakase, H. Suzuki, H. Morinaka, H. Shinohara and K. Mashiko, "An 8.8ns 54 x 54 Bit Multiplier wuth High Speed Redundant Binary Architecture," IEEE Journal of Solid State Circuits, vol. 31, no. 6, pp. 773-783, June 1996 .

[15] J. Fadavi-Ardenkani, "M x N Booth Encoded Multiplier Generator Using Optimized Wallace Trees," IEEE Transaction on Very Large Scale Integration (VLSI) System, vol. 1, no. 2, pp. 120-125, June 1993.

[16] G. DeMicheli and P. Song, "Circuit and Architecture Tradeoffs for High Speed Multiplication," IEEE Journal of Solid State Circuits, vol. 26, no. 9, pp. 1184-1198, September 1991.

[17] L. Dadda, "Some Schemes for Parallel Multipliers," Alta Frequenza, vol. 34, pp. 349-356, 1965.

[18] D. Booth, "A Signed Binary Multiplication Technique," Quarterly Journal of Mechanics and Applied Mathematics, vol. 4, no. 2, pp. 236-240, 1951.

[19] N. Besli, A Novel Arithmetic Unit Using Redundant Binary Signed Digit Number System, Ph.D. Thesis, Florida Institute of Technology, 2004.

[20] "www.altera.com," [Online].

New Simulation Method of New HV Power Supply for Industrial Microwave Generators with N=2 Magnetrons

N.EL GHAZAL, A.BELHAIBA, M.CHRAYGANE, B.BAHANI
Laboratory of Materials, Systems and Information of Technology (MSTI), High School of Technology, Ibn Zohr University, BP: 33/S 80000, Agadir-Morocco

M.FERFRA
Electrical Engineering, Power Electronics Laboratory (EMI), Mohammadia's School of Engineering Mohamed V University, BP: 765, Ibn Sina, Rabat-Morocco

***Abstract*—This original work treats a new simulation method of a new type of high voltage power supply for microwave generators with N magnetrons (treated case: N=2 magnetrons), used as a source of energy in the industrial applications. This new power supply is composed of a single-phase HV transformer with magnetic leakage flow, supplying two parallel cells, which multiplies the voltage and stabilizes the current. The doubler supplies one magnetron. The transformer is presented by its π equivalent circuit. Each inductance of the model is characterized by its relation "flow-current". In this paper, we present a new approach validation of the π model of the special transformer using Matlab-Simulink code. The theoretical results compared with the experimental measurements, is in good agreement with them. The use of this tool Matlab-Simulink, has allowed us to confirm the possibility of the operation of this new system without interaction between magnetrons, with a view to a possible optimization which lead to reduce the weight, the volume and the cost of implementation while ensuring the process of regulating the current in each magnetron.**

***Keywords*—Modeling; New Power Supply; Magnetron; Microwave Generator; Matlab-Simulink; High Voltage (HV)**

I. INTRODUCTION

This work, with industrial character, touches the electrical engineering field for an application hyper frequency linked to the production of the microwave energy. It is within the framework of the development the modeling of manufacturing technologies of power supply for industrial microwave generators and in particular that of his HV transformer with magnetic shunts. This paper discusses a new simulation method modeling a new high-voltage power supply for N=2 magnetrons 800Watts-2450 MHz of the figure 1, used as a source of energy microwave. The design of this new power supply uses a HV transformer with magnetic leakage supplying two cell voltage doublers and stabilizers of current [1-13]. The transformer with magnetic shunts ensures the stabilization of average anode current in each magnetron, thanks to the saturation of his magnetic circuit. The trend towards this new power system will be considered a different version of the single-phase model currently manufactured by the manufacturers of the domestic or industrial microwave ovens.

The modeling of this new generation of power supply for a magnetron passes necessarily by the modeling and the dimensioning of its own new HV transformer with magnetic shunts.

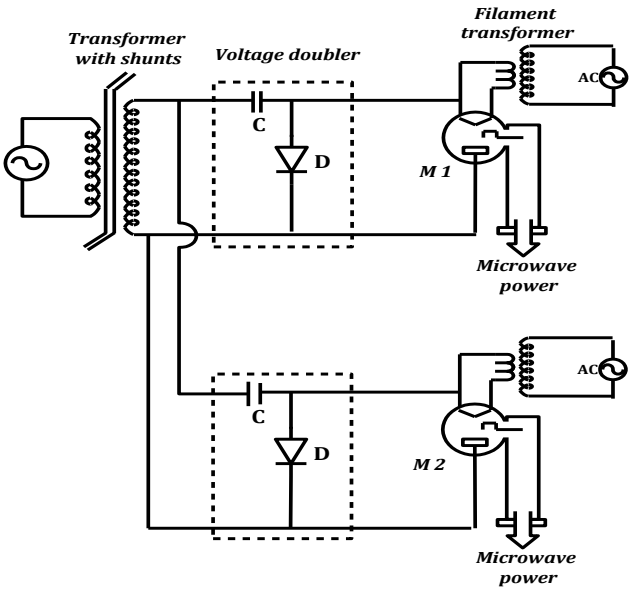


Fig. 1. New single-phase high voltage power supply for N= 2 magnetrons (Amperex Technology)

Our objective in this paper is on the one hand, to treat the modeling of the single-phase power supply designed currently to power normally, in nominal mode, one magnetron which is Moulinex brand. On the other hand, basing on this power supply to treat the modeling, for the first time, of the new power supply for N = 2 magnetrons using the tool Matlab-Simulink. The paper is organized as follows:

Firstly, we discuss the modeling of the single-phase power supply currently used in the microwave generators. The modeling with Matlab-Simulink uses the power supply of the model developed by Mr. Chraygane of the transformer which is a π quadruple. The results will be compared with those obtained experimental.

Secondly, we present a new method validation of the π model of the single-phase HV transformer for microwave generators with N=2 magnetrons. This model has been tested in simulation software digital Matlab-Simulink. The signals obtained will be compared to the experimental measurements.

Thirdly, for the first time, we treat the possibility of the operation of this new system in case of failure of one or two magnetrons. This which allows to obtain relative to the current device gains in space, volume, cost of implementation and maintenance and makes this new device more economical while ensuring the process of controlling the current in each magnetron.

II. Modeling of the Current Power Supply for One Magnetron

The modeling already developed [1-11] of a single-phase HV power supply for one magnetron 800 Watts- 2450 MHz (Figure 2) is to model essentially the special HV transformer with magnetic leakage which ensures the stabilization of the means anodic current in the magnetron.

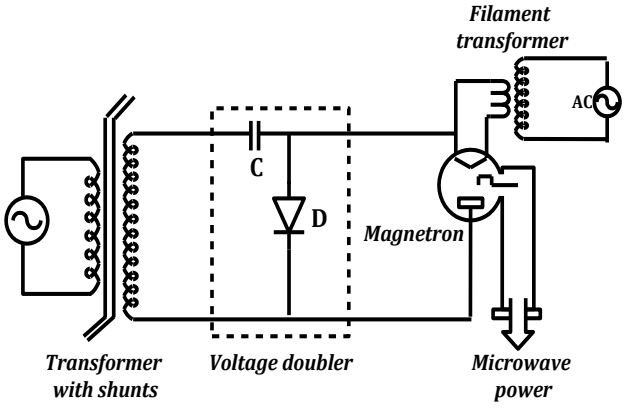


Fig. 2. Current power supply for one magnetron (Amperex technology)

The equivalent model of the selected transformer will be integrated into the overall scheme of the power supply to be adapted to the modeling of the whole system with Matlab-Simulink.

The equivalent circuit should translate the behavior of the whole power including the magnetron and the transformer with shunts. The simultaneous solution of the electric and magnetic equations of the whole system is too complex and the solution can be only digital (Matlab-Simulink) that analytical, with the possibility to study the choice of materials and dimensions of the transformer with possible optimization.

The figure 3 shows the incorporation of the π equivalent circuit of the transformer in the power supply resulting from the electric and magnetic equations of its operation. The advantage of this model is in its single-phase equivalent circuit referred to the secondary, which seems more comfortable to study the operation of the transformer with Matlab-Simulink. This model is called as natural because each inductance, with iron core, is based on the reluctance, therefore the permeability of a very specific part of the magnetic circuit supposed fictitiously closed on which are coiled n_2 turns (secondary). The immediate relevance of this model is to assign at each inductance a non linear relation "flow-current" of the form $n_2\phi(i)$ from the geometrical parameters of a specific portion of the magnetic circuit, allowing translating its real operation in non linear mode.

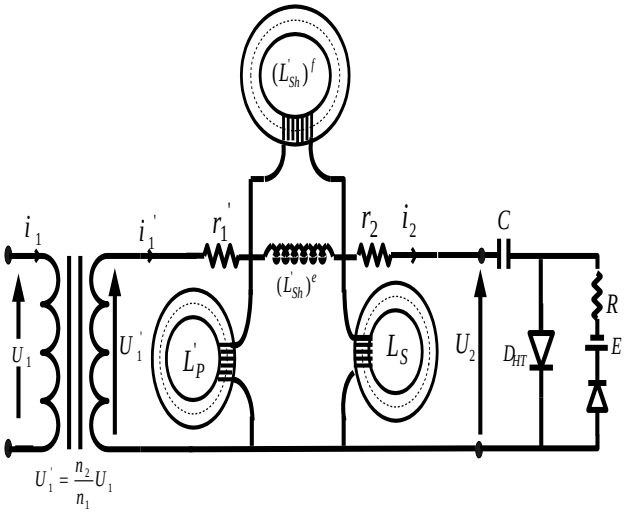


Fig. 3. Π quadruple model of the transformer leakage referred to the secondary

To perform this modeling [1], [3],[4] we have sought to integrate the model of the transformer in the power circuit from the source to the magnetron (Figure 3), where we represented the microwave tube by its equivalent circuit deduced from its electrical characteristic that is formally similar to that of a diode with dynamic resistance of 350Ohms and threshold voltage E≈3800 volts.

The elements of the model, in particular the non linear inductances, were determined from the magnetic characteristics of plates and the geometrical dimensions of the transformer. Each element of a saturable portion of the magnetic circuit, of section ***S*** and of medium length ***l***, is represented by its inductance $L(i)=n2\Phi(i)/i$ where the quantity $n_2\Phi(i)$ and its corresponding current ***i*** can be determined from the curve B(H) of the material used and the geometrical elements using the relations:

$$n_2 * \Phi = n_2 * B * S \;\; \& \;\; i = (H * \ell)/n_2$$

To validate this model, we have carried out tests on a microwave generator composed of the following elements:

- A HV transformer with magnetic shunts characterized by: f=50 Hz, S=1650 VA, U_1=220 V, and U_2=2330 V (resistance of the primary referred to the secondary r'_1=100Ω, secondary resistance r_2=65Ω, number of primary turns: n_1=224, number of turns in the secondary n_2 = 2400).
- A condenser with a capacity C=0.9 μF and a high voltage diode DHT.
- A magnetron designed to function under an approximately voltage ≈4000 V

To obtain its nominal power, it needs an average intensity Imean≈300 mA, but without exceeding the peak value of its current (Ipeak <1,2 A).

In addition, the data from the manufacturer made it possible to extract the values E=3800V and R = 350Ohm.

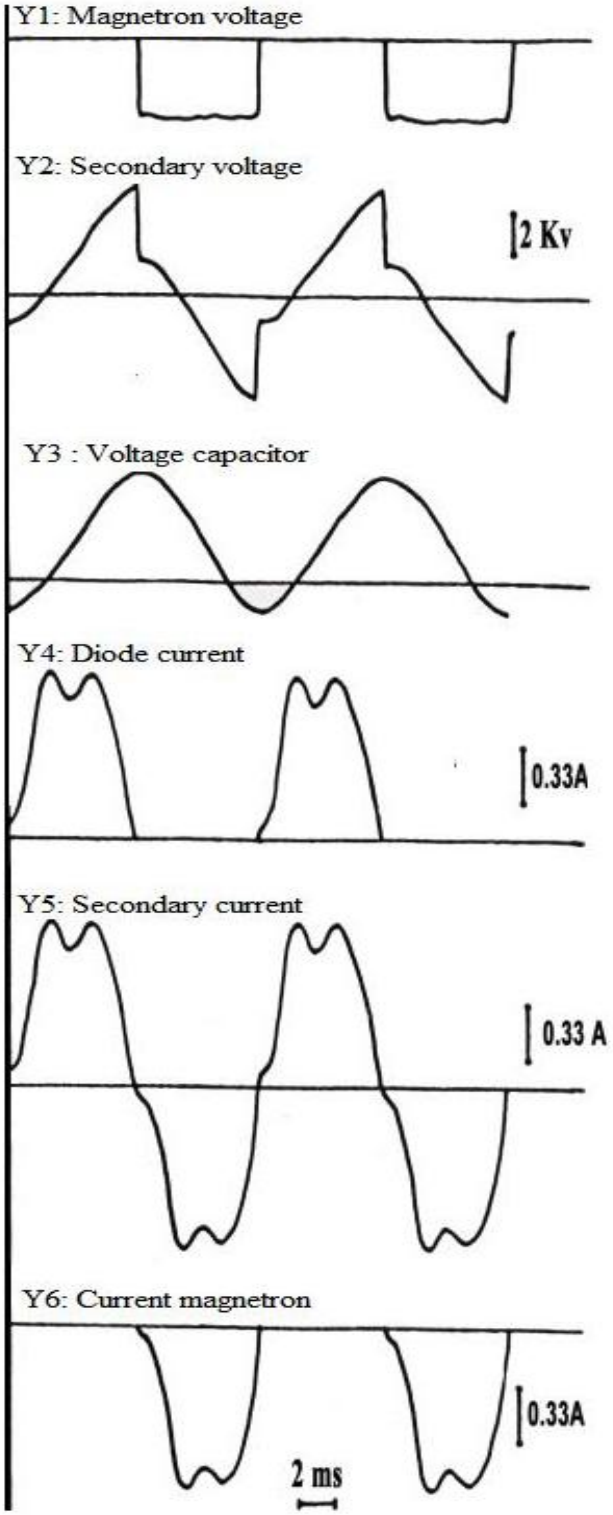


Fig. 4. A. Concordance of the experimental waveforms of currents and voltages (nominal mode)

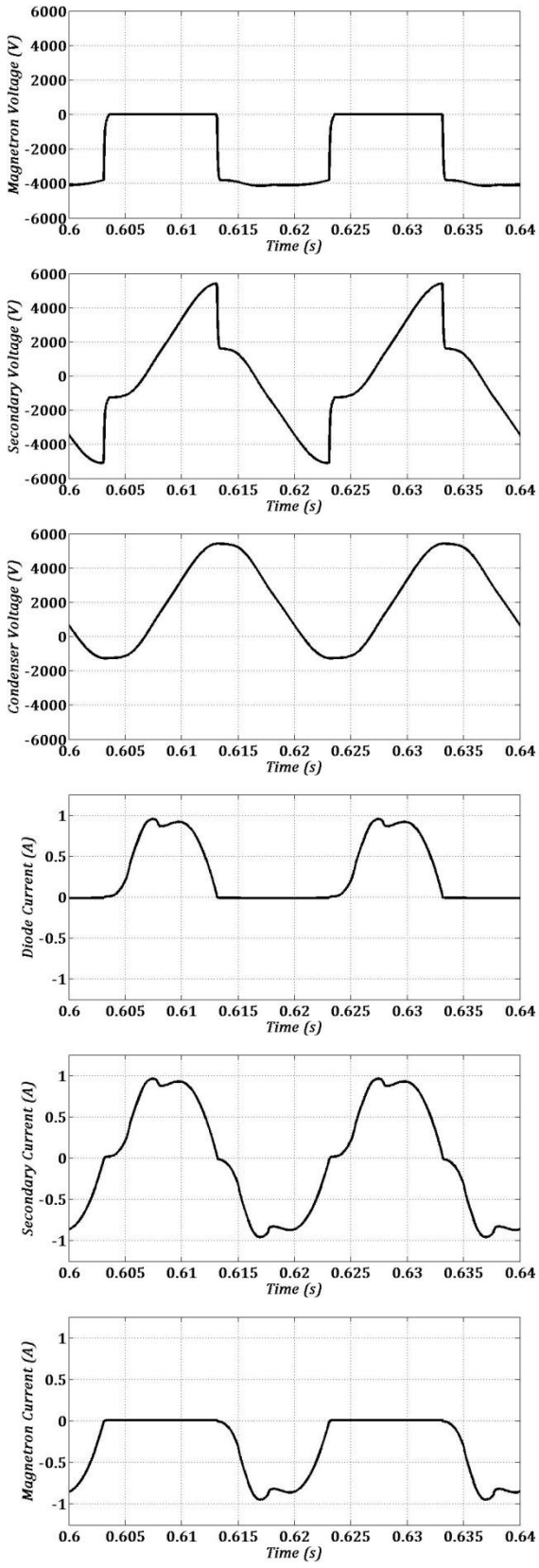


Fig. 4. B. Concordance of the theoretical waveforms, obtained by Matlab-Simulink, of currents and voltages (nominal mode)

The figures 4A and 4B show that in nominal operation (U_1=220 V and f= 50Hz) the results of the simulation by Matlab-Simulink of the device, in non linear regime, are in concordance with the forms of the experimental waves observed in these same conditions. Indeed, between values peak to peak, the relative differences will never exceed 4%.

Taking into account the precision of various data and the acceptable tolerances on operation of the magnetron, the validity of the modeling using Matlab-Simulink was satisfactory. On the other hand, the stabilizing effect of the magnetron current has been verified with respect to the variations of the primary voltage of 10% of the nominal voltage (±20V).

III. Simulation the Model of the New HV Power Supply Under Matlab-Simulink

The modeling previously simulated [13-15] of the new HV power supply for two magnetrons 800 Watts-2450 MHz (Using The EMTP code) is to model essentially its special HV transformer with magnetic leakage. This transformer ensures the stabilization of the average anodic current in each magnetron. Integrated in an overall scheme of the new HV power supply for N = 2 magnetrons, it is suitable for the modeling the whole by a powerful tool for the numerical calculation Matlab-Simulink.

The equivalent diagram must translate the behavior of the whole, including the two magnetrons and the transformer cannot be separated from the external circuits. Too complex to be analytical, the solution can be only digitally using suitable software (Matlab-Simulink).

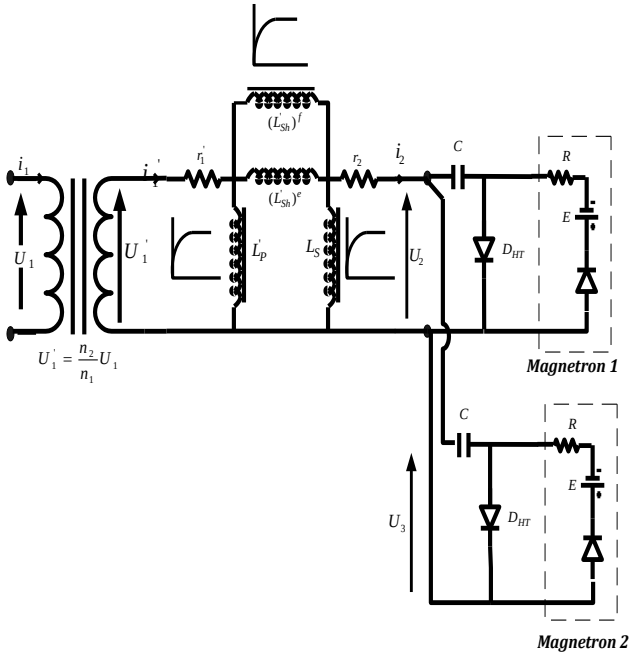


Fig. 5. New power supply for N = 2 magnetrons: Validation of the modeling with Matlab-Simulink

The figure 5 shows the integration of the model of the new transformer in the circuit of the new HV power supply from the source to the magnetrons.

Each inductance of this model is a function of the reluctance of the portion of the magnetic circuit that it represents. The simulation by Matlab-Simulink, in non linear regime linked to the saturation of magnetic circuits, is possible. Then each non linear inductive element is represented by its characteristic, depending on the relation:

$$n_2 * \Phi = n_2 * B * S \;\; \& \;\; i = (H * \ell) / n_2$$

Under matlab-Simulink each inductance is represented by a block diagram (figure 6) and its elements are the following:

- An integrator allows you to deduct the flow from the voltage measurement.
- A function called Loockup table: this is a block which the input-output relation is defined by the user. In the dialog box, which is created when you click on the icon of the LUT, it defines point by point the relation input-output. It is here the couple of values (i,Φ) deducted from those (H,B) and the geometrical data for the three inductances.
- An imposed current source.

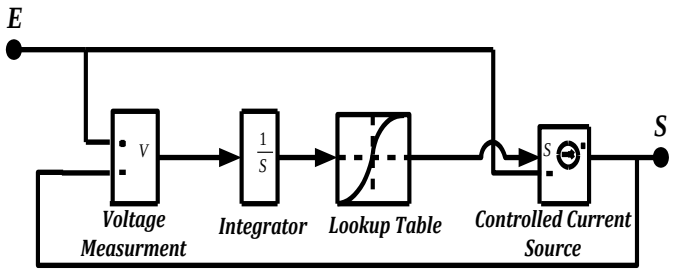


Fig. 6. Block Diagram of a non-linear inductance in Matlab-Simulink.

Thus, the figure 7 shows the overall scheme of the HV power supply for two magnetrons simulated using the software Matlab-Simulink.

We validate this model by comparing the results of simulations with those obtained from the tests already made [13-15] on a generator microwave composed of the following elements:

- A HV transformer with magnetic shunts characterized by: f=50 Hz, S=1650 VA, U_1=220 V, and U_2=2330 V (resistance of the primary referred to the secondary r'_1=100Ω, secondary resistance r_2=65Ω, number of primary turns: n_1=224, number of turns in the secondary n_2 = 2400).
- Two voltage doublers which each one is composed of a condenser with a capacity C=0.9 μF and a high voltage diode DHT.
- Two identical magnetrons which each one is designed to operate under a voltage of about 4000V. For its nominal power, it needs an average intensity I***moy***=300mA, but without exceeding the peak current may destroy it (I***max*** <1.2 A).

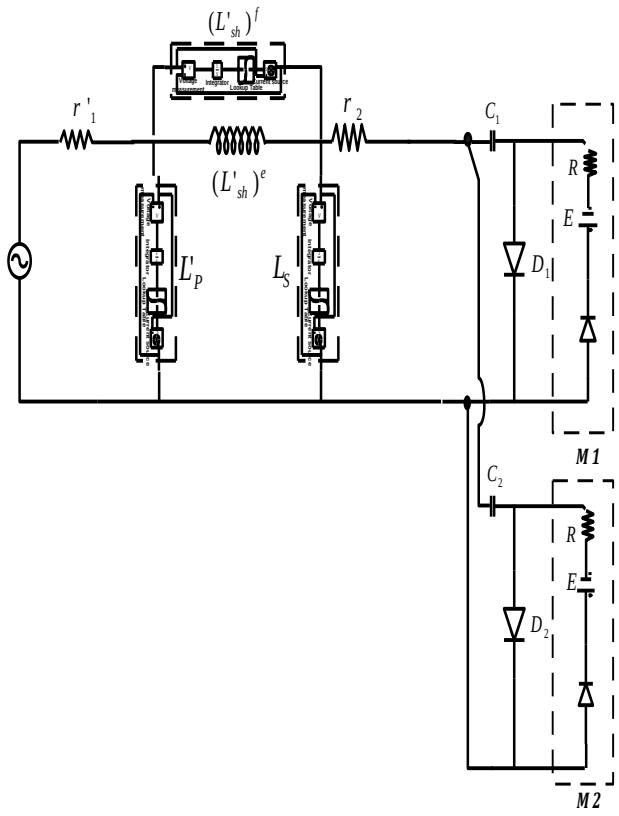


Fig. 7. Circuit of the new HV power supply for N = 2 magnetrons simulated using Matlab-Simulink code in nominal mode (non-linear regime)).

The figures 8 and 9 resulting from simulation of the device by Matlab-Simulink and those obtained in practice are in accordance. The peak current magnetron obtained by Matlab-Simulink (-0.96A) that remains close -1A after practical results. Indeed, from peak to peak value, the relative differences never exceed 6%. The simulation was satisfactory by Matlab-Simulink.

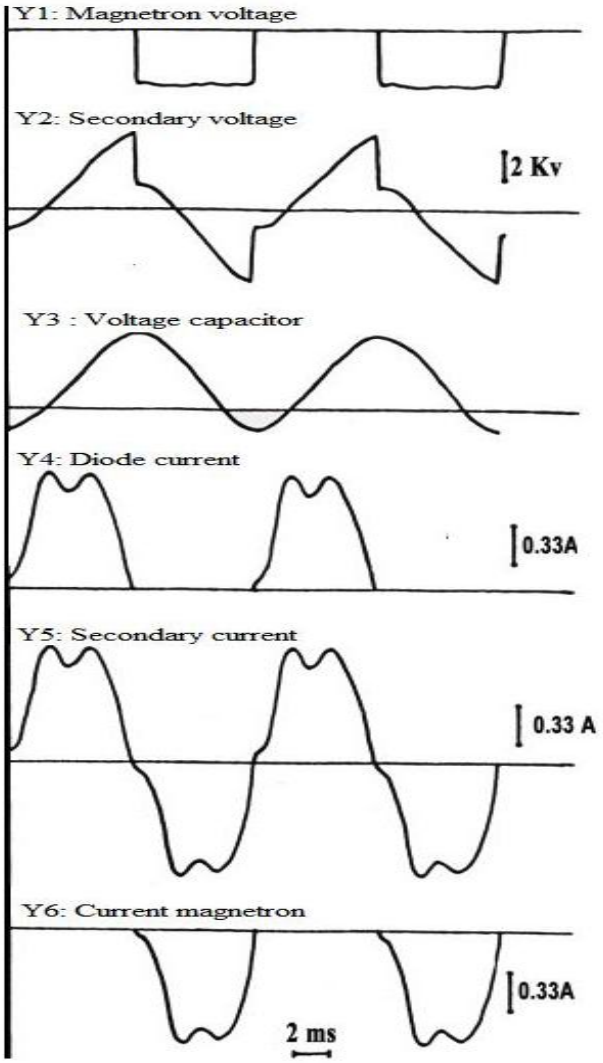


Fig. 8. Experimental waveforms of currents and voltages (nominal mode)

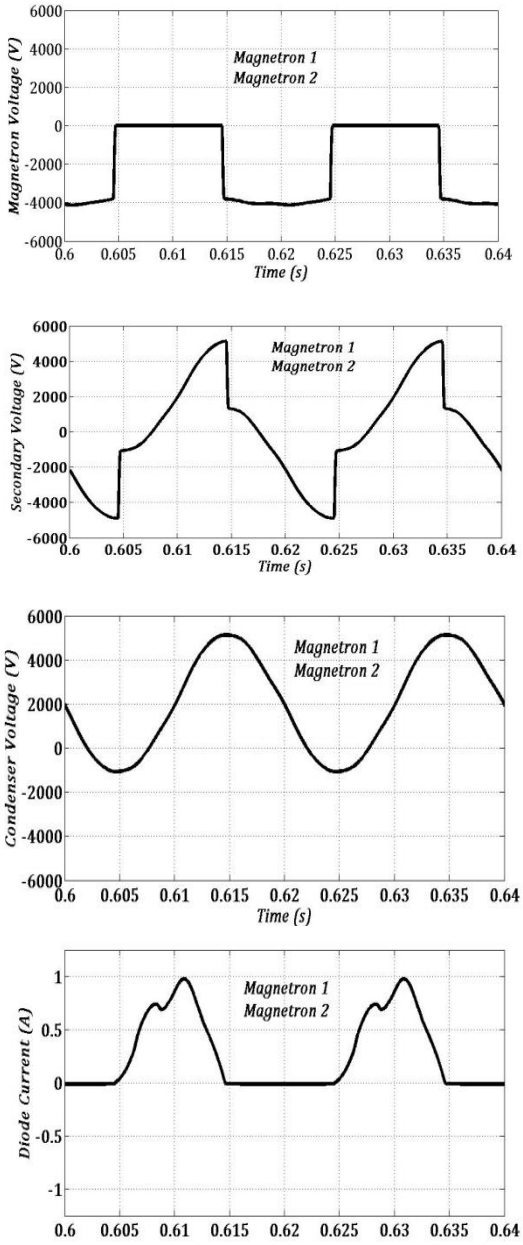


Fig. 9. Simulation with Matlab-Simulink code: waveforms of voltages and currents (nominal mode)

IV. Virification of Functioning of the New Power Supply in Case of Failure

The study of operation of two magnetrons is performed. Now we are going to deal with a study which is the first of its kind in this area, the case where the one of the two or the tow magnetrons fails. We will repeat the simulation and see the curves of currents and voltages using the code Matlab-Simulink to envisage the influence of magnetron failure on the operation of the remaining magnetron.

A. Case of one magnetron in failure

The feasibility study of the operation at the nominal mode of this new system of figure 10 was undertaken. It is to highlight if it is possible to supply a single magnetron which the other is in failure.

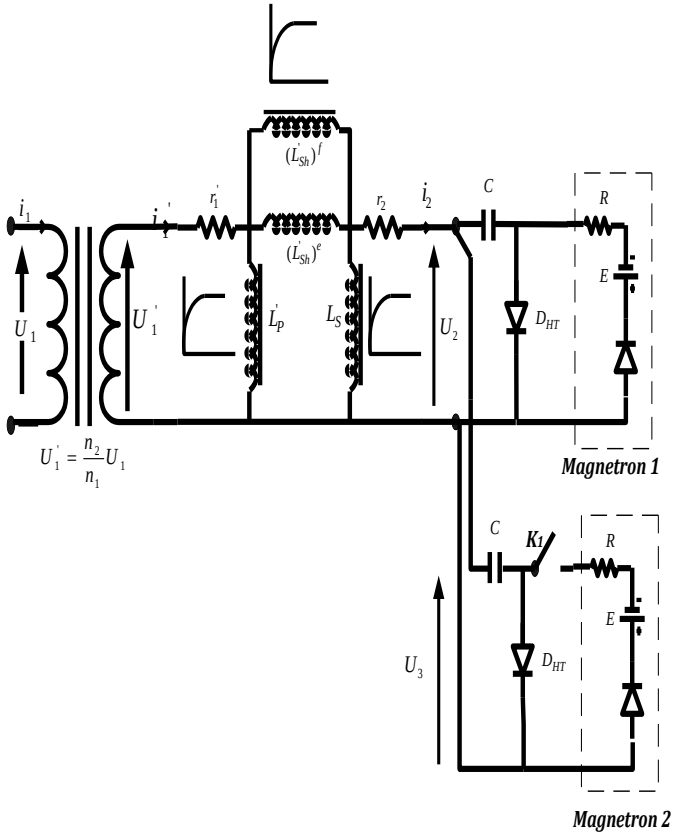


Fig. 10. Simulation of the circuit of the HV power supply using the code Matlab-Simulink in non linear regime for one magnetron on and the other off.

The simulations with Matlab-Simulink the mounting of the figure 10 have helped to raise the temporal oscillograms of the currents and the voltages of the figure 11. The diagram of this figure was simulated to account for the operation of the new power supply for a microwave generator able to deliver, in this case, under 220V the full power 800Watt useful at 2450MHz.

We note that the signals obtained for the magnetron in operation are identical to those in normal operation (without fail) of a conventional power supply with a single magnetron. This confirms the absence of interaction between the magnetrons.

The operating point of the magnetron in operation is therefore not more disturbed, which crucial for a stabilized power supply with current. In addition, the failure of one magnetron does not affect the operation of the remaining magnetron. It suffices to replace the magnetron off by a new magnetron.

B. Case of two magnetrons in failure

Now we are going to treat the case where the two magnetrons are in failure. We are going to repeat the simulation and visualize the curves of currents and voltages using Matlab-Simulink code to envisage the influence of the magnetrons in failure on the normal operation.

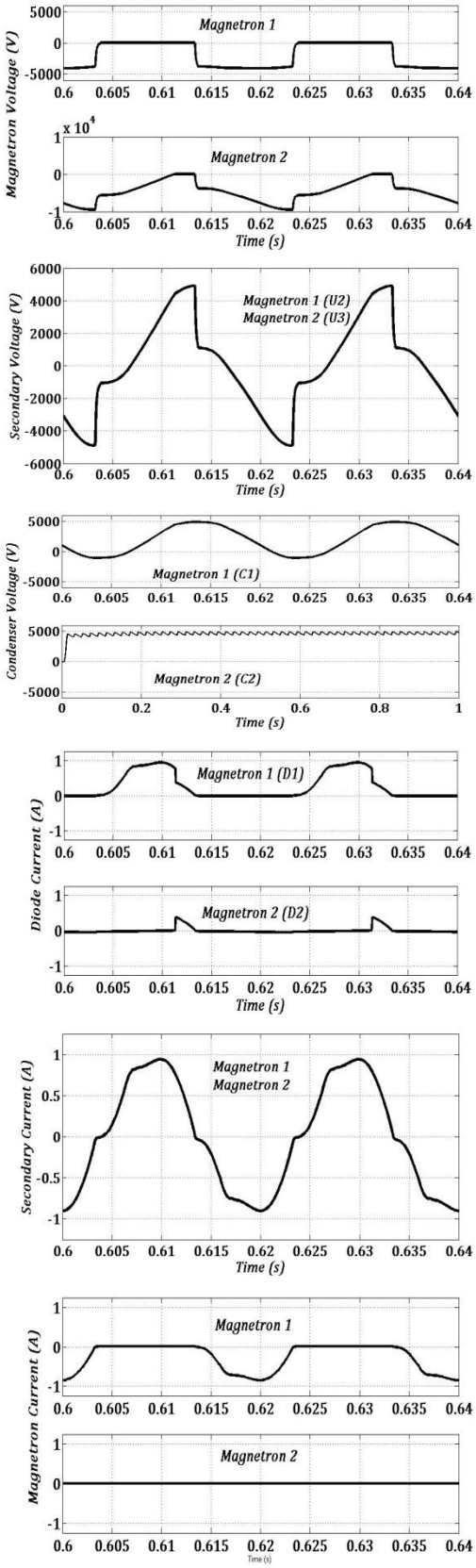


Fig. 11. Oscillograms of currents and voltages of the modeled circuit during the simulation in non linear regime with Matlab-Simulink code (M_2 in failure)

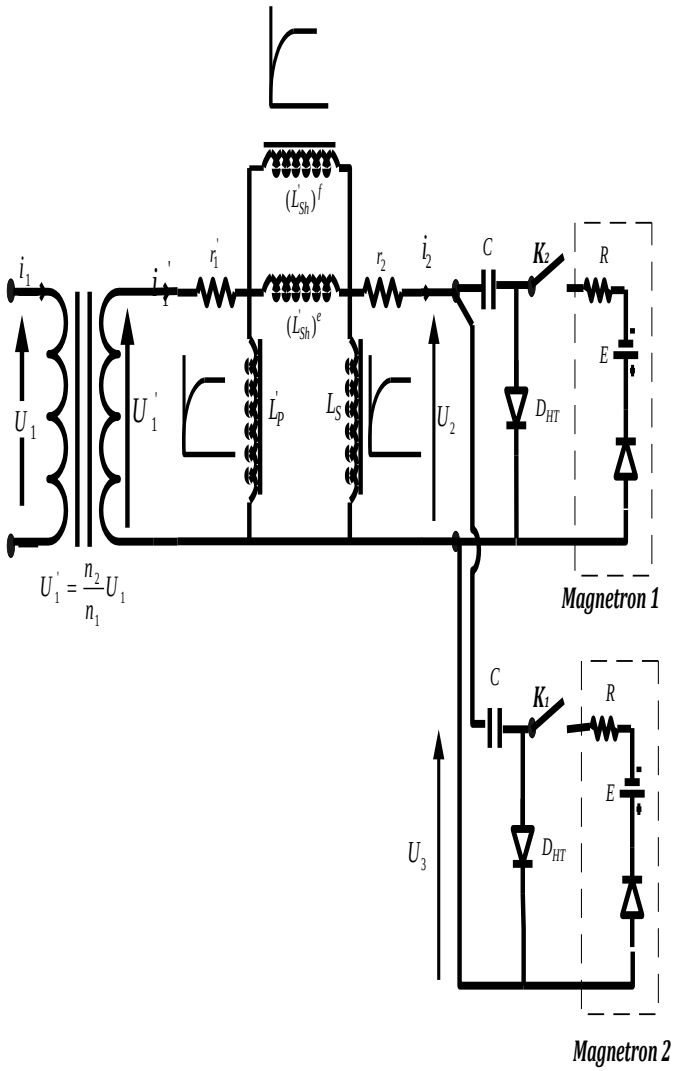
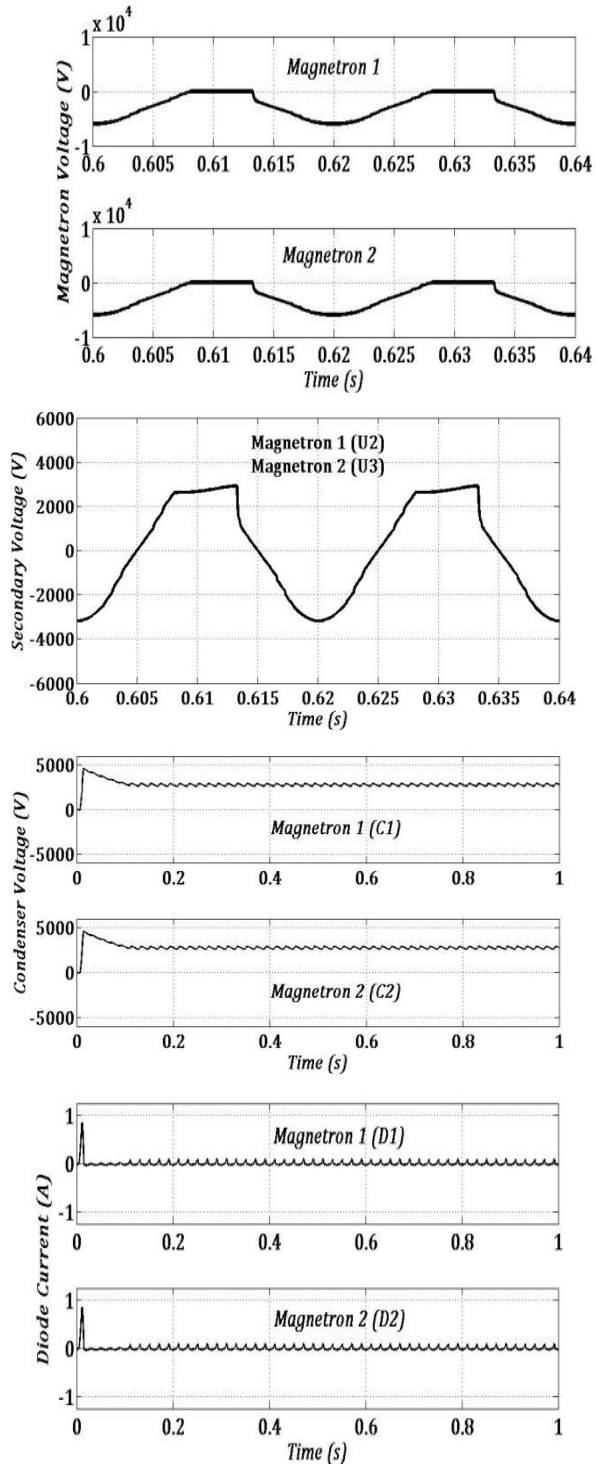


Fig. 12. Simulation circuit of HV power supply using the code Matlab-Simulink in non linear regime for two magnetrons in failure.

Fig. 13. From the non linear characteristics already established of each inductance using Matlab-Simulink code, we simulated the electrical behavior of the HV circuit of the power supply in the Figure 12 where the two magnetrons are off. The oscillograms obtained during this simulation are represented in the Figure 13.

The results in Figure 13 demonstrate that the failure of the tow magnetrons M_1 and M_2 does not disturb the normal operation of the new transformer. So the new system can work without any problems while respecting the constraints recommended by the manufacturer.

The study of operation at the nominal mode of the new power system with two magnetrons in the case of a failure is conclusive. The conclusive study of the new power supply will certainly push us to undertake the study of a new power supply with several magnetron (N >2), which will allow without doubt reduce the volume, the weight and the electrical wiring and therefore guarantee a decrease in the cost of implementation and maintenance of microwave generators.



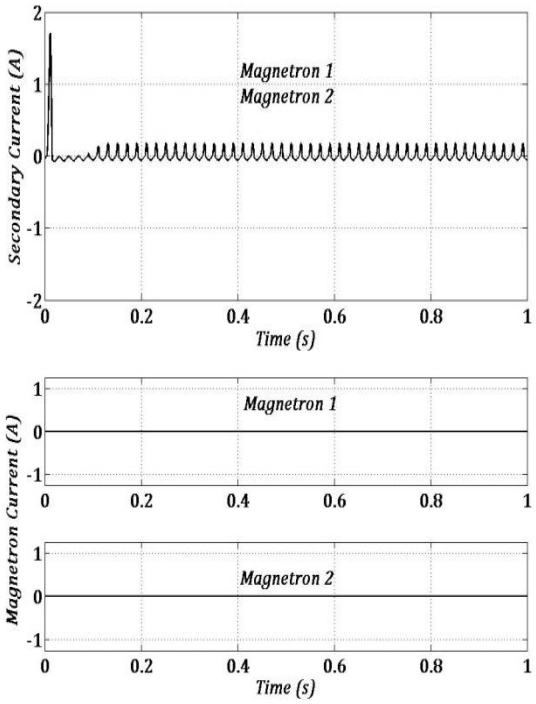


Fig. 14. Oscillograms of currents and voltages of the modeled circuit during the simulation in non-linear regime with Matlab-Simulink code (M_1&M_2 in failure)

V. CONCLUSION

The modeling using Matlab-Simulink code of the nominal operation of a new HV power supply for microwave generators with two magnetrons is conclusive. This new type of power supply can deliver up to 2*800 = 1600 Watts Useful at 2450 MHz after his own adequately sized transformer with magnetic shunts.

The observations made using the Matlab-Simulink code show an excellent agreement between the simulated and the real tests. This code has confirmed the validity of the model in the non linear regime. On the other hand, the failure of one of the two magnetrons not affects the operation of the remaining magnetron.

As perceptive, this work can also be done similarly to the case of the same type of HV power supply for several magnetrons each power output 1000 Watts or 1200 Watts at 2450 MHz, which contributes to the development of the effective modeling of a new systems power for several magnetrons 800 Watts, 1000 Watts or 1200 Watts at 2450 MHz for microwave generators using in industrial applications.

REFERENCES

[1] Chraygane M, Modélisation avec EMTP d'une nouvelle génération d'alimentation haute tension monophasée pour générateurs micro-ondes à magnétrons destinés aux applications industrielles, Thèse de doctorat d'état, Université IBN ZOHR Agadir, Maroc, n° 113/07 (2007)

[2] Dorgelot E.G., Philips Technishe Rundschau, Vol. 21934 (1980) 103-109

[3] Chraygane M, Modélisation et optimisation du transformateur à shunts d'une alimentation haute tension à magnétron pour générateurs micro-ondes 800 W-2450 Mhz destinés aux applications industrielles, Thèse de doctorat, Université Claude Bernard Lyon I, France, n° 189 (1993)

[4] Chraygane M, Teissier M., Jammal A. et Masson J.P Modélisation d'un transformateur à shunts utilisé dans l'alimentation H.T d'un générateurs micro-ondes à magnétron, publication, journal de physique III, France,(1994) 2329-2338

[5] Teissier M., Chraygane M., Jammal A. et Masson J.P , Leakage Flux Transformer Modelling, Communication, International Conference on Electric Machines, ICEM'94, Paris, (1994)

[6] M. Chraygane, M. Ferfra, M. El Khouzaï, B. Hlimi, étude de l'état magnétique interne global du transformateur à shunts d'une alimentation pour générateurs micro-ondes à magnétron destinés aux applications industrielles, Télécom'2003 et 3ème JFMMA–Marrakech. Comm., (2003) 436-439.

[7] M. Chraygane, M. Ferfra, M. El Khouzaï, B. Hlimi, Vérification expérimentale de la loi de conservation des flux du transformateur à shunts d'une alimentation pour générateurs micro-ondes à magnétron destinés aux applications industrielles, RNJCP4-Casablanca. Comm., (2003) 6-7.

[8] Chraygane, M. Ferfra, B. Hlimi, Modélisation d'une alimentation haute tension pour générateurs micro-ondes industriels à magnétron, Revue 3EI, 41 (2005) 37-47.

[9] M. Chraygane, M. El Khouzaï, M. Ferfra, & B. Hlimi, Etude analytique de la répartition des flux dans le transformateur à shunts d'une alimentation haute tension pour magnétron 800 Watts à 2450 Mhz, J. of PCN, 22 (2005) 65-74.

[10] M. Chraygane, M. Ferfra, & B. Hlimi, Etude analytique et expérimentale des flux du transformateur à shunts d'une alimentation pour magnétron 800 Watts à 2450 Mhz, J. of PCN, 27, (2006) 31-42.

[11] M. Chraygane, M. Ferfra, B. Hlimi, Détermination analytique des flux et des courants du transformateur à fuites d'une alimentation haute tension à magnétron pour générateurs micro-ondes industriels 800 Watts à 2450 Mhz, J. of PCN, 40 (2008) 51-61.

[12] Aguili T & Chraygane M., Une alimentation originale pour générateurs micro-ondes, Revue Générale de l'Electricité RGE n° 5, France, (1990) 49-51.

[13] M. Chraygane, A. Zatni, M. Ferfra, B.Hlimi & S. Bidar, Modélisation d'une nouvelle alimentation HT monophasée pour générateurs micro-ondes industriels à N=2 magnétrons, Télécom'2007 et 5ème JFMMA–Fès. Comm., (2007) 420-424.

[14] M. Chraygane, M. Ferfra, M.El Haziti, A.Zatni, M. Bour, M. Lharch, Modélisation et simulation du fonctionnement nominal d'une nouvelle alimentation HT monophasée pour générateurs micro-ondes industriels à N=3 magnétrons, communication, Télécom'2009 et 6ème JFMMA-Agadir. Comm., (2009) 77-78.

[15] M. Ferfra, M. Chraygane, M. Fadel, M. Ould Ahmedou, Modélisation non linéaire d'une nouvelle alimentation haute tension globale de N=2 magnétrons pour générateurs micro-ondes industriels. [Non linear modelling of an overall new high voltage power supply for N=2 magnetrons for industrial microwave generators], J. of PCN, 54 (2010) 17-30.

[16] Hermann W. Dommel, ElectroMagnetic Transcients Program, Reference Manual, EMTP Theory Book, 1986

[17] Van Dommelen D., ATP General Introduction, Leuven EMTP Summer Course, July 1991.

[18] Minutes of the 20 th European EMTP Users Group Meeting, Leuven EMTP Center, October 28-29 th, 1991.

[19] W. Scott Meyer et Tsu-huei Liu, Alternative Transients Program (ATP), Rule Book, Canadian/american EMTP User group, 1987-92.

[20] Laurent Dubé, European EMTP-ATP Users Group e.V, Users Guide to models in ATP, April 1996.

[21] Hans K. Hoidalen, Atpdraw for Windows, Atpdraw Version 3 User Manuel (Atpdraw Installation Manual, Atpdraw Introductory Manual, Atpdraw advanced Manuel), European EMTP-ATP Users Group e.V, February 1996.

[22] Mustafa Kizilcay et Laszlo Prikler, European EMTP-ATP Users Group e.V, EEUG News, Number 3, Volume 3, August 1997.

[23] Mustafa Kizilcay et Laszlo Prikler, ATP-EMTP Beginner's Guide for EEUG Members, European EMTP-ATP Users Groupe.V, June 2000.

[24] Emperreur G., Transformers modelling basic theory, exemples, Leuven EMTP Summer Course, Belgium, July 1991.

[25] Roguin J., Ranjamina V., Modeling of magnetic circuits with EMTP, EDF, bulletin de la DER – série B, réseaux électriques, matériels électriques, N°2, pp 23-26, 1986

[26] Capolino G. A., Simulation for powers electronics and drivers using ATP, Leuven EMTP summer course, July 1991.

AUTHORS PROFILE

NAAMA EL GHAZAL was born in Laayoune, Morocco, in 08/04/1984; he received the Master Instrumentation and Telecommunications in 2010 from the faculty of sciences (Ibn Zohr University) Agadir-Morocco, where he pursues his doctoral program. His research is interested in the '' Feasibility study in nominal operation of a new three-phase high voltage power supply for industrial microwave generators with N magnetrons per phase''.

ABDERRAHIM BELHAIBA was born in Agadir, Morocco, in 06/12/1983; he received the Master in 2010 in instrumentation and Telecommunications from faculty of sciences (Ibn Zohr University) Agadir-Morocco, where he pursues his doctoral program. His research is interested in the '' Study of the energy balance of a new high voltage power supply N magnetrons per phase 800 Watts - 2450 MHz for microwave generators used in industrial applications''.

DR. MOHAMMED CHRAYGANE was born in Morocco in 1963; he received his thesis of doctorat from Claude Bernard University Lyon I in 1993 and his 'doctorat d'état' from Ibn Zohr University Agadir-Morocco in 2007. In 1994, he joined Technology Higher School Ibn Zohr University Agadir Morocco

(ESTA). Since this date he has been a professor in MSTI Laboratory (ESTA School Ibn Zohr University Agadir Morocco Agadir). His field of interest is modeling a high voltage power supply used for industrial microwaves generators with magnetron.

DR. MOHAMMED FERFRA was born in Rabat Morocco in 1965, he received the engineering degree from Mohammadia's School of Engineering (Mohamed V University) Rabat-Morocco in 1988. From 1988 to 1990, he was an assistant in the same school. From 1990 – 1993, he pursued his PhD program at Laval University, Quebec Canada, where he received PhD degree in Electrical Engineering in 1993. Since this date he has been a professor with the department of electrical engineering at EMI. His field of interest is system identification of electrical machines and modeling shunt transformer used for microwaves generators.

BOUBKER BAHANI was born in Rabat, Morocco, in 15/06/1986; he received the Master in 2010 in Instrumentation and Telecommunications from the faculty of sciences (Ibn Zohr University) Agadir-Morocco, where he pursues his doctoral program. His research is interested in the '' Modeling with EMTP and Matlab of new generation of overall HV power supply for N magnetron 800 Watts-2450MHz for industrial microwave generators ''.

A New Image-Based Model For Predicting Cracks In Sewer Pipes

Iraky Khalifa
Computer Science Department
Faculty of Computers and Information Helwan University
Cairo, Egypt

Amal Elsayed Aboutabl
Computer Science Department
Faculty of Computers and Information Helwan University
Cairo, Egypt

Gamal Sayed AbdelAziz Barakat
Holding Company for Water and Waste IT Department
Egypt

***Abstract*—Visual inspection by a human operator has been mostly used up till now to detect cracks in sewer pipes. In this paper, we address the problem of automated detection of such cracks. We propose a model which detects crack fractures that may occur in weak areas of a network of pipes. The model also predicts the level of dangerousness of the detected cracks among five crack levels. We evaluate our results by comparing them with those obtained by using the Canny algorithm. The accuracy percentage of this model exceeds 90% and outperforms other approaches.**

Keywords—Visual inspection; Sewer pipes; Canny algorithm; Crack detection

I. INTRODUCTION

There is an urgent need to develop a proactive sewer pipeline crack prediction model. Due to the fact that sewer pipelines are hidden from day to day view, deterioration can occur unnoticed and this can in turn lead to unexpected functional failures. Moreover, maintenance and rehabilitation of aging sewers have become an overload in terms of budget allocation and investment planning for towns [1]. There are many factors that may lead to deterioration in the condition of sewer pipelines. These factors may be related to the physical structure of the sewer pipelines such as length, diameter, material and depth. Nevertheless, factors contributing to such deterioration may also be environment-related such as the type of soil and waste [2,3].

There has been a number of studies on the interpretation of sewer pipes inspection data for the purpose of detecting cracks. Moselhi et al [5] described image analysis and pattern recognition techniques of sewer inspection, based on neural network analysis of digitized video images. The neural network analysis technique was found helpful in identifying four categories of sewer defects: cracks, joint displacements, reduction of cross-sectional area. Chae at al [6] developed an automated sewer inspection data interpretation system.

The other approach is to predict a sewer's existing condition prior to its detailed inspection for selective, cost effective sewer inspection. Hasegawa et al. [7] developed a method for predicting sewer pipes condition based on the knowledge of pipe material, length, diameter and other characteristics.

However, it was concluded that the method could not evaluate sewer's condition effectively. Ariaratnam et al [4] developed a logistic regression model for condition evaluation of sewers. The model was developed through historical data based upon factors; such as, pipe age, diameter, material, waste type and depth. Another approach for condition assessment of large sewers was developed by assessing the impact of different factors; such as location, size, burial depth, functionality etc. in [8]. Baur et al [9] developed a methodology of forecasting condition of sewers by using transition curves.

These transition curves were developed through the historical condition assessment data. Sewers characteristics; such as, material, period of construction, location were used to define the existing condition of sewers for scheduling detailed inspection. Yan et al [10] proposed a fuzzy set theory based approach for a pipe's condition assessment. Various linguistic factors: soil condition, surroundings, etc, were transformed through fuzzy theory into numerical format for assessing their impacts on pipes. Ruwanpura et al [1] used rule-based simulation methodology to predict condition rating of sewers.

The model predicted the condition rating of pipe based on age, material and length of pipe. Najafi et al [11] developed an artificial neural network model for predicting the condition of sewers based on historical data. The above mentioned approaches tend to predict existing condition of sewers for prioritizing detailed inspections. Table I shows the comparison between these 7 models including the strong and the weak points.

This paper proposes a model for predicting cracks in sewer pipes cracks. In section II, the data set including sewer pipes image collection is described. Section III presents crack detection using canny algorithm while section IV presents our proposed model for crack detection and crack dangerousness level prediction. Experimental work and results are presented in section V followed by a conclusion in section VI.

II. SEWER PIPES IMAGE COLLECTION

A commercially available SONY-DSC T5 digital camera of 5.1 mega pixels with optical zoom 3x has been used for data collection of 101 crack surface defects. Fig. 1 shows one of these images. These images have been taken to Cairo sewer pipes network. MATLAB functions DILATE (), THRESH () and LAPLACIAN () were developed and applied to this image collection.

TABLE I. COMPARISON BETWEEN THE 7 PREVIOUS MODELS

Author	Strong Points	Weak points
Moselhi [5]	Giving more details of cracks and joint.	Disability to predict the developments of cracks.
Chae [6]	Using Sewer Scanner and Evaluation Technology (SSET)	The used camera is very old so the results are weak.
Hasegawa [7]	Using material, length, diameter and other characteristics for predictions.	Could not evaluate sewer's condition effectively.
Ariaratnam [4]	The model was developed through historical data based upon factors; such as, pipe age, diameter, material, waste type and depth.	Disability to find the relationship between all parameter.
Baur [9]	Developing a methodology of forecasting the condition of sewers by using transition curves.	Depends on historical assessment data only.
Ruwanpura [1]	Using rule-based simulation methodology to predict condition rating of sewer pipes based on age, material and length of pipe.	Makes the task of planning, prioritizing and allocating funds a complex exercise
Najafi [11]	Using artificial neural network model for predicting the condition of sewers based on historical data.	The way of obtaining historical data is unclear.

Fig. 1. One of the 101 trail images

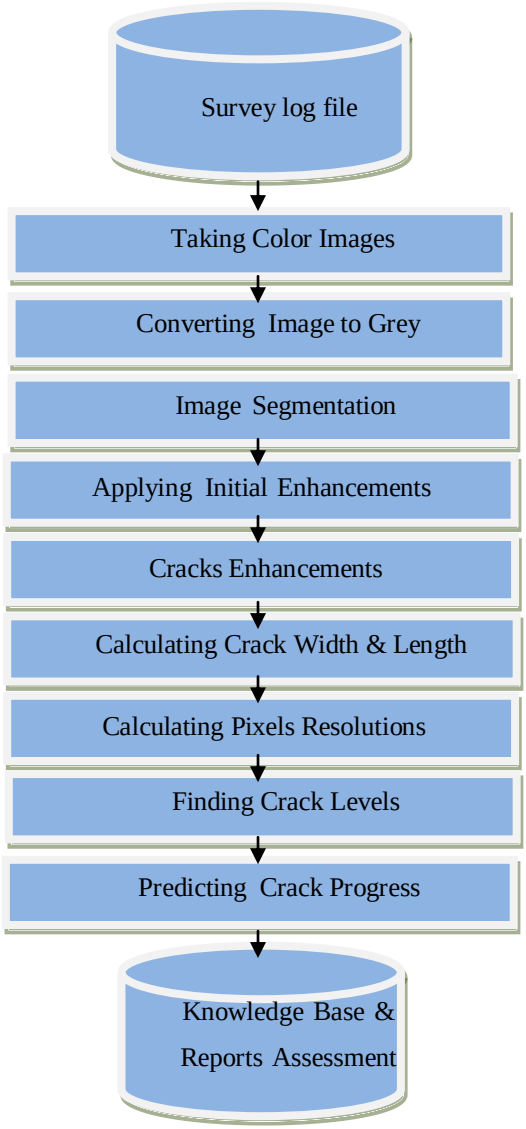


Fig. 2. Steps of the crack detection model

III. CRACK DETECTION USING CANNY EDGE DETECTOR

Segmentation of an image entails the division or separation of the image into regions of similar attribute. The most basic attribute for segmentation is image luminance amplitude for a monochrome image and color components for a color image. Image edges and texture are also useful attributes for segmentation [15].

Image segmentation techniques include thresholding methods, boundary/edge-based methods and region based methods. This section presents the steps needed to implement the Canny edge detector for crack detection.

The first step is to filter out any noise in the original image before trying to locate and detect any edges. After smoothing the image and eliminating noise, the second step is to find the edge strength by taking the gradient of the image. The Sobel operator performs a 2-D spatial gradient measurement on an image. Then, the approximate absolute gradient magnitude (edge strength) at each point can be found. The Sobel operator uses a pair of 3x3 convolution masks Fig. 3, one estimating the gradient in the x-direction (columns) and the other estimating the gradient in the y-direction (rows). These masks are applied to every pixel in the image.

-1	0	+1
-2	0	+2
-1	0	+1

G_x

+1	+2	+1
0	0	0
-1	-2	-1

G_y

Fig. 3. The 3x3 convolution masks used by the Sobel operator

The third step is to find the edge direction which is computed simply using the formula:-

$$\Theta = \tan^{-1}(G_y / G_x). \quad (1)$$

Once the edge direction is known, the next step is to relate the edge direction to a direction that can be traced in an image. So if the pixels of a 5x5 image are aligned as follows:

x x x x x
x x x x x
x x a x x
x x x x x
x x x x x

It can be observed by looking at pixel "a" that there are only four possible directions when describing the surrounding pixels - 0 degrees (in the horizontal direction), 45 degrees (along the positive diagonal), 90 degrees (in the vertical direction), or 135 degrees (along the negative diagonal). Now, the edge orientation has to be resolved into one of these four directions depending on which direction it is closest to (e.g. if the orientation angle is found to be 3 degrees, make it zero degrees).according to equation (1) .

IV. Crack Detection Model

We propose a model to discover sewer pipes cracks. A series of steps have to be performed Fig. 2.

A. Converting Colored Images to Grey

The proposed system has been used for binary (black and white) images and has been extended later to be used with grayscale images as well. The light and dark portions of an image can be reshaped or morphed in various ways under a control of a structuring element which can be considered as a parameter to morphological operation [12]. The MATLAB function im2bw() is used to convert a colored image to a binary image .

B. Image Segmentation

Segmentation of an image entails dividing or partitioning of an image into regions of similar attributes. The most basic attribute for segmentation is image luminance amplitude for a monochrome image and color components for a colored image.

Image edges and texture are also useful attributes for segmentation [13].

Sets A and B in z (image) are defined to represent a grey-level image consisting of pixels p(x, y) and a structuring element (Fig. 4), respectively:

$$A=\{(x,y)|p(x,y)\} \quad (2)$$

$$B=\{(x,y)|p(x,y)\} \quad (3)$$

Converting scanned images into binary images for segmenting pipe defects by using a thresholding method and determining the optimal thresholds for the opening operated gray-level images by maximizing the following measure of class separability [17]:

$$D(T) = \frac{P_1(T)P_2(T)[m_1(T) - m_2(T)]^2}{P_1(T)\sigma_1^2(T) + P_2(T)\sigma_2^2(T)} \quad (4)$$

Where

$$P_1(T) = \sum_{z=0}^{T} h(z) \quad (5)$$

$$P_2(T) = \sum_{z=T+1}^{L-1} h(z) = 1 - P_1(T) \quad (6)$$

$$m_1(T) = \frac{1}{P_1(T)} \sum_{z=0}^{T} z h(z) \quad (7)$$

$$m_2(T) = \frac{1}{P_2(T)} \sum_{z=T+1}^{L-1} z h(z) \quad (8)$$

$$\sigma_1(T) = \frac{1}{P_1(T)} \sum_{z=0}^{T} [z - m_1(T)]^2 h(z) \quad (9)$$

$$\sigma_2(T) = \frac{1}{P_2(T)} \sum_{z=T+1}^{L-1} [z - m_2(T)]^2 h(z) \quad (10)$$

Equations (4) to (9) apply segmentation in the original images . z is the grey-level of a pixel in the scanned image and ranges from 0 through L - 1, h(z) is the normalized grey-level histogram of the scanned image. By maximizing the criterion function in Eq. (3), the means of the light and dark image regions can be separated as much as possible and the variances of the two image regions can be minimized.

C. Applying Initial Enhancements on Images

The contrast of the RGB pipe image has to be improved by enhancing the dark (crack) pixels relative to the background image. In order to perform crack enhancements, erosion and dilation operators are used. Erosion and dilation are two fundamental operators of digital image processing and whose implementations are of great value in the area of image analysis.

The dilation process is performed by laying the structuring element B on the image A and sliding it across the image .

The dilation of A by B, $A \oplus B$, is the union of all pixels in A surrounded by the shape of B [14] and is defined as:

$$A \oplus B = \{a + b \quad \forall a \in A \text{ and } b \in B\} \quad (11)$$

The dilation algorithm is applied as:

1) If the origin of the structuring element coincides with a 'white' pixel in the image, there is no change; move to the next pixel.

2) If the origin of the structuring element coincides with a 'black' in the image, make black all pixels from the image covered by the structuring element. Fig. 5 shows the shape of the structure element.

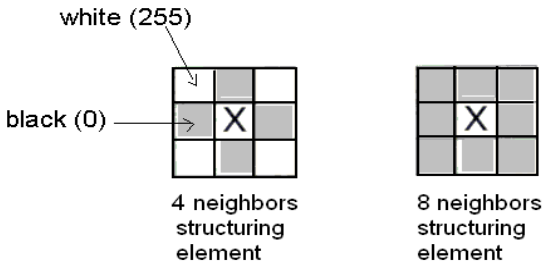


Fig. 4. Fig. 4. Typical shapes of the structuring elements (B)

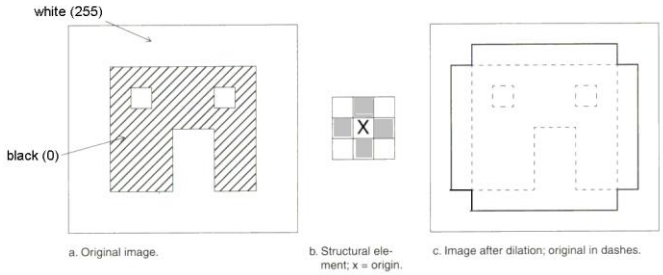


Fig. 5. Illustration of the dilatation process

The example shown in Fig. 5 shows that with a dilation operation, all the 'black' pixels in the original image will be retained, any boundaries will be expanded, and small holes will be filled.

The erosion process is similar to dilation, but we turn pixels to 'white', not 'black'. As before, slide the structuring element across the image and then follow these steps:

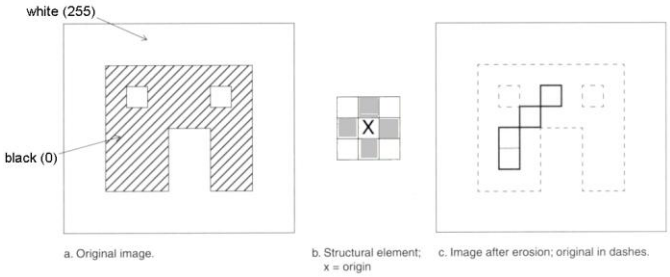


Fig. 6. Illustration of the erosion process

1) If the origin of the structuring element coincides with a 'white' pixel in the image, there is no change; move to the next pixel.

2) If the origin of the structuring element coincides with a 'black' pixel in the image, and at least one of the 'black' pixels in the structuring element falls over a white pixel in the image, then change the 'black' pixel in the image (corresponding to the position on which the center of the structuring element falls) from 'black' to a 'white'.

Erosion of A by B, denoted as $A \Theta B$, removes all pixels within a distance B from the edge of A (Fig. 6) and is defined as:

$$A \Theta B = \{a | b + a \in A \text{ for every } b \in B\} \quad (12)$$

The opening operation is defined as :

$$A \circ B = (A \Theta B) \oplus B \quad (13)$$

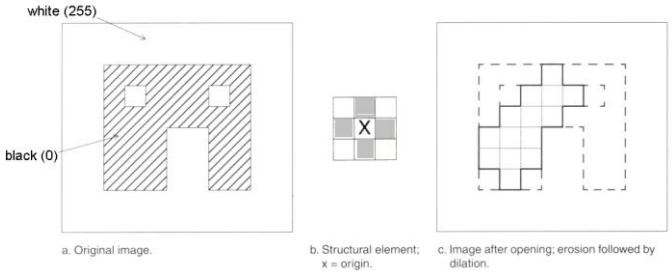
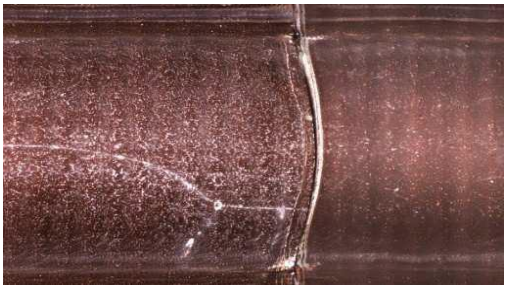


Fig. 7. Illustration of the opening process

These two basic operations, dilation and erosion, can be combined into more complex sequences as shown in Fig 7. The most useful of these for morphological filtering are called opening and closing [19]. Opening consists of an erosion followed by a dilation and can be used to eliminate all pixels in regions that are too small to contain the structuring element. In this case the structuring element is often called a probe, because it is probing the image looking for small objects to filter out of the image.

The effect of opening operation is to remove image regions which are lightly relative to the structuring element while preserving image regions greater than structuring elements [16].

Converting scanned images into binary images for segmenting pipe defects by using a thresholding method and, determining the optimal thresholds for the opening operated gray-level images by maximizing the following measure of class separability [17]:



(a)Original image with pipe crack

(b) Image after applying dilation with min=0 and max=255

Fig. 8. Pipe crack image before and after dilation

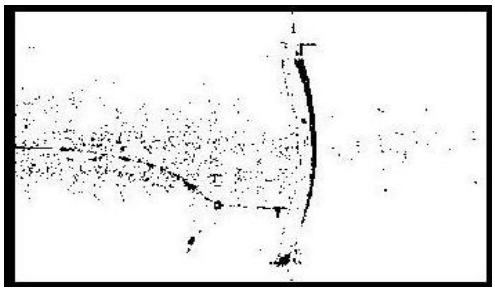


Fig. 9. Applying filters to find cracks

D. Crack Enhancement

Before finally detecting the cracks, some enhancement steps are applied to the preprocessed image as shown in Fig. 8 and Fig. 9. The first enhancement operation on the image is applying a Laplacian filter. The Laplacian of an image f(x,y) denoted $\nabla^2 f(x,y)$ is defined as [18] :

$$\nabla^2 f(x,y) = \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2} \quad (14)$$

Commonly used digital approximations of the second derivatives are

$$\frac{\partial^2 f}{\partial x^2} = f(x+1,y) + f(x-1,y) + 2f(x,y) \quad (15)$$

and

$$\frac{\partial^2 f}{\partial y^2} = f(x,y+1) + f(x,y-1) + 2f(x,y) \quad (16)$$

From equations (14), (15) and (16), it is deduced that

$$\nabla^2 f = f(x+1,y) + f(x-1,y) + f(x,y+1) + f(x,y-1) + 4f(x,y) \quad (17)$$

This expression can be implemented at all points (x, y) in an image by convolving the image with the following spatial mask:

$$\begin{matrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{matrix}$$

An alternate definition of the digital second derivatives takes into account diagonal elements, and can be implemented using the mask:

$$\begin{matrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{matrix}$$

Both derivatives sometimes are defined with the signs opposite to those shown here, resulting in masks that are the negatives of the preceding two masks. Enhancement using the Laplacian is based on equation (18).

$$g(x,y) = f(x,y) + c[\nabla^2 f(x,y)] \quad (18)$$

where $f(x,y)$ is the input image, $g(x,y)$ is the enhanced image, and c is 1 if the center coefficient of the mask is positive, or -1 if it is negative [18] Because the Laplacian is a derivative operator, it sharpens the crack but drives constant areas to zero. Adding the original image back restores the gray-level color .

E. Calculating Crack Width and Length

In the process of calculating the crack width, it is not needed to divide the width into segments because the maximum width measured is 43.132 mm. On the other hand, since the crack length may reach 910.876 mm, it is necessary to divide the total length into segments when the crack length is calculated. Each segment doesn't exceed 100 mm in length which helps to improve the results.

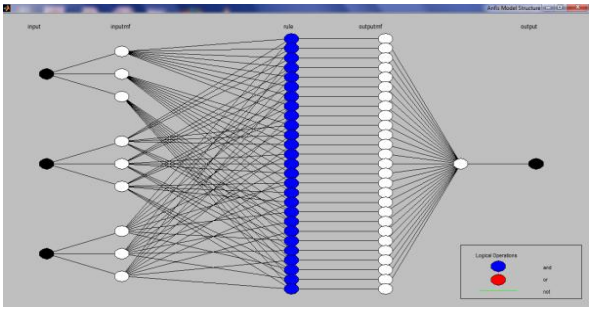


Fig. 10. The inputs (crack width, crack height and pixels) and output crack levels

F. Calculating Pixel Resolution

Resolution refers to the number of pixels in an image and describes the extent of details that can be appreciated in an image. The image resolution and the surface area can be related to determine the quantity of space that each pixel represents in the trail image (Fig. 11). Equations (19) and (20) are used to determine the pixel space represented by an image. Pixel space is represented by the pixel width and height denoted as C_w and C_h respectively. Image trail width and height are denoted by w and h respectively while c is the number of pixel columns and r the number of pixel rows in the image.

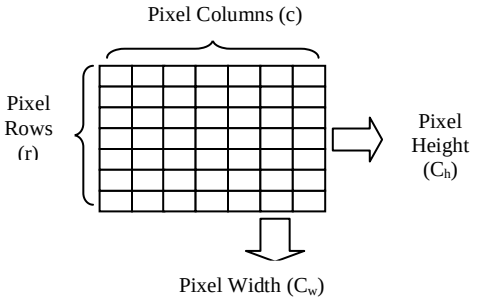


Fig. 11. A part of a digital image showing pixel columns and rows and also pixel height and width

$$C_w = \frac{w}{c} \quad (19)$$

$$C_h = \frac{h}{r} \quad (20)$$

Cracks have less numbers of pixels as compared to their background. PC represents the pixels in cracks while PB represents the pixels in background.

the maximum pixels in cracks in the trail imaged are not greater then 30 PPI as shown in equation (19) .

$$\sum_{n}^{1} PC < \sum_{n}^{1} PB \qquad (21)$$

Measuring crack widths depend on many parameters such as:-

S : no of segments (the default no of segments is >= 1)

V: average of crack width in one image

P_s: The total pixel in one segment

C_p: The total pixel in one crack

n: The number of segments in one crack

K_w: The total pixel in segment width

$$K_w = \frac{1}{n}\sum_{s=1}^{n} w_s \qquad (22)$$

$$C_p = \frac{1}{n}\sum_{s=1}^{n} p_s \qquad (23)$$

G. Finding Crack Levels

Using fuzzy logic, three inputs (crack width, crack height and number of pixels) are used to classify the levels of crack as shown in Fig. 10. The output represents the crack level.

About 80% of the image data set is used as training data and about 20% is used for testing.

$$C_l = C_w/C_p \begin{Bmatrix} .1 \geq C_l < 3 \\ 3 \geq C_l < 6 \\ 6 \geq C_l < 9 \\ 9 \geq C_l < 12 \\ C_l \geq 12 \end{Bmatrix} \qquad (24)$$

TABLE II. FIVE LEVELS OF CRACKS

Degree	C_l	No. of images	C_l %	C_l levels
Very high	5	2	1.98%	$C_{l5}>12$
High	4	2	1.98%	$9 \geq C_{l4} < 12$
Intermediate	3	9	8.91%	$6 \geq C_{l3} < 9$
Low	2	28	27.72%	$3 \geq C_{l2} < 6$
Very low	1	60	59.41%	$0.1 \geq C_{l1} < 3$

V. EXPRIMENTAL WORK AND RESULTS

The proposed model can predict five levels of sewer pipes image cracks as shown in equation (24) and in Table II. Crack level 5 represents the most dangerous level where crack level C_l is greater than 12. The total number of pixels in each crack is not greater than 20 but the total number of pixels in each background in not less than 100. Fig.13 shows the pixels average is between 2 and 14. Using equations (20-24), pipe cracks which cannot be seen by the human eye can be discovered. Then, the crack level is predicted as explained previously. Fuzzy logic in MATLAB and Image J ver1.46r application are used for our implementation. The error percentage is calculated by comparing testing and training data as shown in Fig. 12.

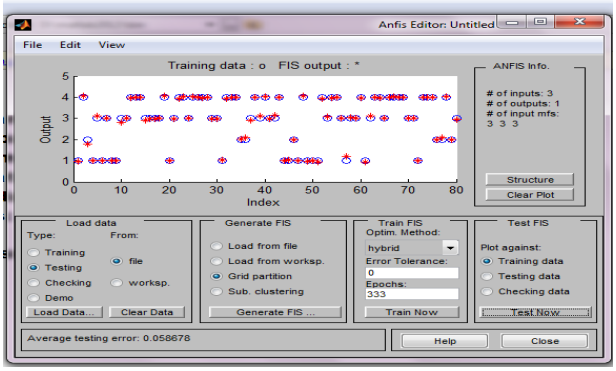


Fig. 12. Average testing and average training error where training data is blue and testing data is red

TABLE III. PROPOSED MODEL VS CANNY DETECTION

Name	Crack detection%	Crack Prediction %
Proposed Model	91	90
Canny Detection	86	23

TABLE IV. ACCURACY AND ERROR PERCENTAGES IN IMAGE SAMPLES

Degree of dangerousness	No. of Images	Accuracy Results
Very high	2/2	100%
High	2/2	100%
Intermediate	9/8	89%
Low	28/24	86%
Very Low	60/54	90%

The very high and high levels of cracks dangerousness are the first priority to the decision maker. Such cracks are very difficult to detect by the human eye as the number of pixels in a crack is smaller than the number of pixels in the image background. Table IV shows the accuracy percentage obtained for each crack level.

The accuracy of the proposed model is greater than or equal to 90% for crack detection and crack prediction. The accuracy of Canny model is 86% in crack detection and very low in crack prediction as shown in Table III .

This is due to the fact that edge directions according to the Canny algorithm are four directions only (0, 45, 90, 135 degrees). Any edge direction falling within the range (0 to 22.5 and 157.5 to 180 degrees) is set to 0 degrees. Any edge direction falling in the range (22.5 to 67.5 degrees) is set to 45 degrees. Any edge direction falling in the range (67.5 to 112.5 degrees) is set to 90 degrees. And finally, any edge direction falling within the range (112.5 to 157.5 degrees) is set to 135 degrees. Moreover, removing the noise in the Canny algorithm affects the image structure leading to an increase in the error percentage.

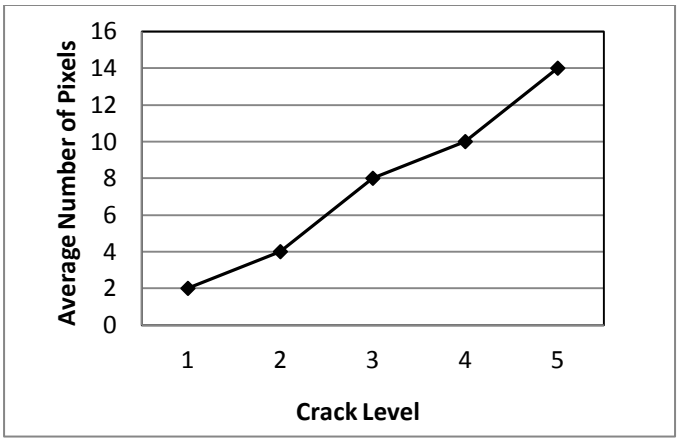


Fig. 13. The relationship between crack levels and average number of pixels

As shown in Table V, the proposed model has many advantages over the previous models. The accuracy of the proposed model is about 90% which is greater than the accuracy of the previous models. The proposed model has 5 crack levels but the other models have only one crack level. The availability of a number of levels enables the operator to make a more accurate decision. The proposed model focuses on the crack so it helps to predict what will happen to this crack in the future.

TABLE V. COMPARISON BETWEEN THE PROPOSED MODEL AND THE 7 PREVIOUS MODELS

Items	Proposed model	Pervious 7 model
Accuracy of prediction	About 90%	About 60%
Number of Crack levels	5	1
Attributes included	Crack width, crack height, crack pixels	Focus on conditions and other materials only
Cost	Need very cheap tools	Need very expensive machines
Easiness	Very easy	Very difficult because of using complicated machine
Main technique	Taking many regular images	Used tools may miss defects hidden behind obstructions or under water

VI. CONCLUSION

Most crack detection techniques depend on the human eye. This paper presents an analytical model together with its implementation for the purpose of detecting cracks as well as predicting their level of dangerousness. Using our model, very small cracks which can't be detected by the human eye can be detected and their level of dangerousness can be predicted. Five levels of dangerousness can be predicted. The accuracy of both detection and prediction exceeds 90% which outperforms other approaches.

REFERENCES

[1] Ruwanpura J, Ariaratnam, S, and El-Assaly, A, (2004), "Prediction Models for Sewer Infrastructure Utilizing Rule-Based Simulation", Journal of Civil Engineering and Environmental Systems, volume 21, No 3, Page 169-185

[2] F. Chughtai and T. Zayed, "Infrastructure Condition Prediction Models for Sustainable Sewer Pipelines", Journal of Performance of Constructed Facilities, vol. 22(5), October 2008

[3] Rahman, S & Vanier D, (2004), "An Evaluation of Condition Assessment Protocols for Sewer Management", National Research Council of Canada Research Report Number B-5123.6

[4] Ariaratnam, S, El-Assaly, A & Yang, Y, (2001), "Assessment of Infrastructure Inspection Needs using Logistic Models", ASCE Journal of Infrastructure Systems, Volume 7, No 4, December 2001

[5] Moselhi, O and Shehab-Eldeen, T, (2000), "Classification of Defects in Sewer Pipes using Neural Networks", ASCE Journal of Infrastructures System, Volume 06, Number 03, September, 2000

[6] Chae, M and Abraham, M, (2001), "Neuro-Fuzzy Approaches for Sanitary Sewer Pipeline Condition Assessment", ASCE Journal of Computing in Civil Engineering, Volume 15, Number 1, January, 2001

[7] Hasegawa, K, Wada, Y & Miura, H, (1999), "New Assessment System for Premeditated Management and Maintenance of Sewer Pipe Networks", Proceedings of 8th International Conference on Urban Storm Drainage, Page 586-593, Sydney, Australia

[8] McDonald, S & Zhao, J, (2001), "Condition Assessment and Rehabilitation of Large Sewers", Proceedings of International Conference on Underground Infrastructure Research, page 361-369, Waterloo, Canada

[9] Baur, R & Herz, R, (2002), "Selective Inspection Planning with Ageing Forecast for Sewer Types", International Water Association (IWA) Journal of Water Science and Technology, Volume 46, No 6-7, page 389-396

[10] Yan J & Vairavamoorthy, K, (2003), "Fuzzy Approach for Pipe Condition Assessment", Proceedings of the American Society of Civil Engineers (ASCE) International Pipeline Conference, USA

[11] Najafi M & Kulandaivel G, (2005), "Pipeline Condition Prediction Using Neural Network Models", Proceedings of the American Society of Civil Engineers (ASCE) International Pipeline Conference, USA

[12] Sinha, S and Fieguth, P, (2006), "Segmentation of sewer Concrete Pipe Images", Journal of Automation in Construction, Volume 15, Pages 47-57

[13] Dingus, M., Haven, J., and Russell, A. _2002_. Nondestructive, noninvasive assessment of underground pipelines, AWWA Research Foundation, Denver

[14] Y Dong, BC Forster, AK Milne, Comparison of radar image segmentation by Gaussian- and Gamma-Markov random field models. Int. J. Remote Sens. 24(4), 711–722 (2003). doi:10.1080/0143116021000013322

[15] PIKS Scientific Inside, WILLIAM K. PRATT ,DIGITAL IMAGE PROCESSING, Los Altos, California, A John Wiley & Sons, Inc,2007

[16] Sinha, S. K., & Fieguth, P. W. (2006). Segmentation of buried concrete pipe images. Automation in Construction, 15(1), 47–57.

[17] Yan, H. (1996). Unified formulation of a class of image thresholding techniques. Pattern Recognition, 29(12), 2025–2032

[18] Gonzalez and Woods Prentice Hall ,Errata and Clarifications Digital Image Processing 3rd Edition © 2008 September 10, 2008 CORRECTIONS

[19] . Umbaugh Scot E, Computer Vision and Image Processing, Prentice Hall, NJ, 1998, ISBN 0-13-264599-8

The cybercrime process : an overview of scientific challenges and methods

Patrick Lallement
Charles Delaunay Institute,
University of Technology of Troyes (UTT)
12 rue Marie Curie, CS 42060, 10004, Troyes Cedex, France

***Abstract*—The aim of this article is to describe the cybercrime process and to identify all issues that appear at the different steps, between the detection of incident to the final report that must be exploitable for a judge. It is to identify at all steps, issues and methods to address them.**

Keywords— cybercrime; detection; forensic analysis

I. INTRODUCTION

The cyber criminality is generaly not defined as a whole science, neither a field (in France dor ex. there is no laboratory devoted to this transversal domain). The cybercrime field is generally viewed as an application domain for many communities concerned by information or data processing, decision-making aid, detection methods, sociology, networking, etc. That is why main issues are generally addressed in a fragmented way. However, the forensic process whose aim is to collect and process digital evidences raises different issues because systems are more and more complex and criminal strategies are continuously changing.

This paper approaches this topic by the way of the process that investigators use after an incident. Part I of the paper describes the different tasks and actions that constitute this process. They are two key-words to define the cybercrime process: the detection (of something abnormal) and the investigation for digital evidence (digital forensic). Part III presents different aspects of detection challenges: intrusion (in a system), fraud, suspicious contents. Part IV is dedicated to investigations (how to collect and qualify relevant data). Part V presents the challenge concerning forensic analysis: how to organize digital evidence and organize a pertinent argumentation. Part VI makes a general synthesis of all these new challenges.

II. THE CYBERCRIME PROCESS

A. Definition of Cybercrime

The cybercrime is defined in the penal law as a set of malicious acts that are committed against information systems or that make use of information and communication technologies (ICT). In the first subset we can class denial of services (DoS) attacks, theft or falsification of data. policy. This detection function can be executed in reactive mode as control function of a system or in a proactive mode by law enforcement actions such as internet flow or social networks supervision to look for suspicious contents. In the first case, the abnormal event (or state) is falsification of data.

The second subset concerns fraud, child pornography, sexual harassment by the way of internet and all logistic support activities of organized criminalities.

B. Cybercrime vs Security

The cybercrime process is initiated when the detection function identifies a situation or event as abnormal referring to the assumed security level and the security detected by processing control variables; in the second case, the nature of application contents carried though networks may alert about an illegal activities. Fig. 1 shows out the cybercrime process regarding to the security process in the case of any information system.

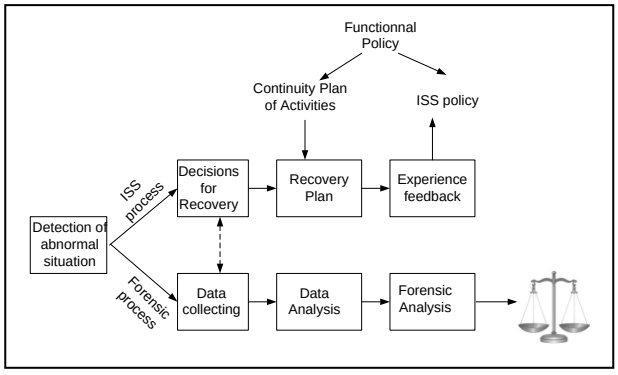


Fig. 1. Cybercrime process vs. security process

The detection function can be considered as common to both processes although the case of APT malwares (Advanced Persistent Threats), where investigations and system recovery take a long time and should be processed commonly and as to avoid alerting the intruding malware itself. The detection function must detect an abnormal situation and qualify it as malicious or not. If then, it generates an alerting event. Fig.2 presents the different functions that will succeed to the detection and take place in the cybercrime process: investigation (collecting of clues, qualification of evidences), forensic analysis, argumentation and final reporting. The digital evidence has to be built from data collected on the (cyber) crime scene [1]. The digital evidence must show out a link between an attacker and a victim [2]. As consequence of their digital aspects, they may be heterogeneous, altered, uncertain and corrupted [3]. They have to be analyzed, interpreted and documented by forensic examiners such as they can be reliable and relevant to draw their conclusions for a court.

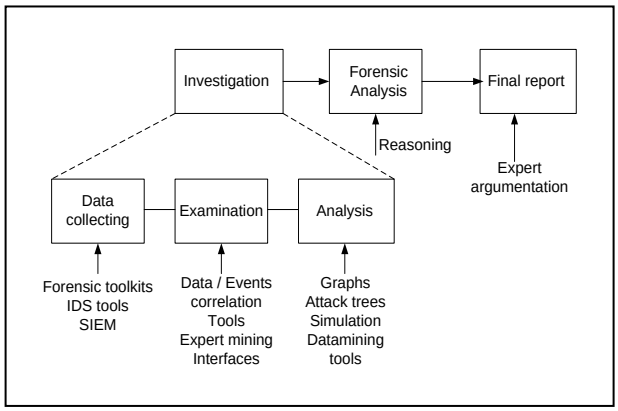


Fig. 2. Cybercrime process

III. DETECTION

Detection comes within the competence of forensic experts, but generally, it cannot be performed humanly, due to the complexity of the system under control, the volume of information to process, the velocity of some attacks, their critical aspects and the predefined scenario that they can use to run. An automatic detection system is then relevant to cope with many of these challenges to react with a more or less expert way, a more or less proactive scheme; they can only generate burglar alarms (less) or be able to characterize and counter the malicious event (more).

A. Intrusion detection

It is made with Intrusion Detection Systems (IDS) [4]. The IDS objective is to detect an abnormal state and to qualify it as intrusive state or not, and if so, to trigger an alert. IDS can report alert in IDMEF format [5] based on XML syntax, which can be useful to organize a cooperative surveillance for distributed systems, and perform alerts correlations. This is of greater importance as countermeasure against some attacks which runs according to a pre-defined scenario. In this case, an IDS-level detection can be completed with a real-time layer to predict what is happening and prevent some severe attacks (distributed DoS, rootkits, worms) that can spread using a slow and/or sophisticated propagation scheme (botnets, rootkits). This layer developed for IPS (Intrusion Prevention Systems) requires reasoning methods at a global level as for IPS. They use generally Bayesian approaches and variables from the network. Bayesian networks offer a powerful way for modeling, representing and reasoning with complex information and have been proposed to process alert correlations systems [6]. For large systems, this reasoning layer can represent a fast mining challenge, using complex time-stamped events [7].

Detection can also be performed in a signal mode, using statistical approaches that present the advantage to avoid a priori knowledge (comparing with Bayesian approaches) [8]. The variables used are: the traffic rate, abnormal packets, CPU utilization, etc. A likelihood function can be built and the challenge is to minimize the false positive and the false negative ratio.

Concerning intrusion detection, the remaining issue lies in the description of complex situations; a language has been proposed by the ANR PLACID project (2007-2011) with the Intrusion Detection Description Logic (IDDl) [9] to describe intrusions, it is IDMEF-compatible, but it is limited to handle information about alerts, topology and vulnerabilities.

B. Fraud detection

The detection challenge can be assimilated as a classification problem between legitimate and fraudulent transactions. The methods used can be supervised or not. In supervised mode, models request learning to distinguish between legitimate and fraudulent transactions. Because fraudulent ones are less frequent (< 1%) than the others, they are worse learned and therefore, the classification quality is decreased. Artificial neural networks have been largely proposed in the 90s, Support Vector Machine (SVM) [10] [11] but their efficiency is largely depending on the type of transaction considered. In [12] authors have compared the different classification methods among various applications. More recent works have suggested a fusion approach with different methods, to filter the current transactions with a level of suspicion, to use the Dempster-Shafer theory to quantify an overall belief for a transaction, to use history and a Bayesian learner to classify suspicious transactions [13].

In non-supervised mode the learner doesn't use any a priori class. It must be designed to the specific context: insurances, payment, telecoms… Methods proposed are based on graphs, decision trees, neural networks, fuzzy rules.

Some works suggest a combination of supervised and non-supervised approaches. In addition to the unit fraud detection problem, a correlation between them may be necessary to identify organized group frauds. The more recent techniques aim to integrate business rules and social networks data.

C. Suspicious content detection : steganalysis

It makes reference to data dissimulated behind a legal flow (voice, video). Detection of hidden data remains a difficult issue because it exploits opportunities given by the coding techniques. Detection methods in signal mode have been proposed [14] [15].

D. Detection n peer to peer networks (P2P)

Content analysis in network to detect malicious activities will concern the internet in general but more precisely social networks and P2P exchanges. In this case, the detection problem becomes rather an identification problem (of paedophile activity for ex.). Data to examine are of a huge volume, they are also dynamic and in the case of P2P networks, there is no central authority. A random and not computer-aided flow inspection is not possible because a large amount of data is necessary to build an evidence of illegal activities. Moreover and at the difference to the previously mentioned detection methods, no history is there available because illegal behaviours always try to be undetectable by using encrypting tools or specific key-words. Neither learning nor statistical methods are efficient here. Approaches proposed are rather inference based on expert (law enforcement)-defined rules to detect and process queries [16]. IP addresses are relied to UDP flows to identify the users. Nevertheless, the computer-aided and automated tools to state that a given user is for ex. a paedophile stands a legal problem, the expert should always

have the last word and automated tools should be viewed as processing resource to cope with the large amount of data.

E. Detection in social networks

The detection issue is there doubled: it is to detect communities on micro-blogging platforms and then to detect specific breaches in violation of citizen protection (fraud, illegal content dissemination, attack to underage, etc.). To help law enforcement people, a processing chain must then associate, in detection and investigation modes, the content analysis of publications and conversations and also the analysis of relations between actors, while it could capture knowledge about the structure, the behaviour and the practices of criminals. Social networks pose complex issues concerning contents and network analysis and also a visualization challenge. It is due to the large amount of data to process (for ex. 465 Millions of tweeter accounts, 175 Millions of tweets per day), the velocity (< 1 minute) and the variety of data (structured and not-structured).

Annotation of texts from social networks is difficult, due to the flow processing and to the downgraded linguistic nature of messages and conversations, which are also multi-languages and multi-domains. Approaches used are rather symbolic, statistic, but the most promising seems to be the mix of them.

The networks analysis uses graph-based representations. There is no consensus to describe and quantify the dynamic of graphs, and to describe how the information does propagate along them. Several works have studied how to retrieve comprehensive information from the structure of static cyber-communities from complex networks [17]. The identification issue for dynamic communities is now addressed by two ways: 1) a dynamic graph can be viewed as a succession of static graphs, each of them representing a state of the dynamic graph at a given moment. In each static graph (i.e. at each time) it is the possible to determine communities with more or less independencies. It is then necessary to retrieve correspondences between communities along the time to restore the temporal evolution. More complex rules have to be defined to identify fusion, scission, appearance and extinction of communities [18] [19] [20]. 2) Specific algorithms have to be designed to detect communities in dynamic graphs [21] [22].

The social networks analysis methods are borrowed from the graph theory and are completed with many works about data and text-mining to process data extracted from social networks messages and relations, indicators and aggregates computed from social graphs and the dynamics of exchanges. Techniques developed recently from the "pervasive computing" domain give interesting perspectives [23]. In this frame, social networks users are viewed as "sensors" that give information about its environment. New sensors can then enhance the already existing sensors. The more recent works suggest to use data-fusion techniques, Complex Event Processing (CEP) engines, time-sequences association and analysis, spatiotemporal patterns to detect events (alarm reporting, weak signals).

F. Synthesis

In all cases, the detection issues have to cope with a large amount of heterogeneous data. In some cases, there are also serious time constraints. Most of methods proposed are similar to those for decision-making. Indeed, the reasoning associated to detection consist in doing classification between normal, abnormal and suspicious cases, and then to decide if the suspicious case is normal or not, using supplementary data such as history, learning techniques and quantitative methods developed in the artificial intelligence field (fuzzy rules, neural networks). Most of them are used to detect intrusion or frauds. Methods that are designed with a generic approach are rare [24], probably because specific information (contexts, experience, behaviours) is necessary to reduce false positive and false negative ratios.

IV. INVESTIGATIONS

The response to incident process can be split in several steps: data-gathering, examination, analysis, reporting [25] [26]. Data are of different volatility as defined in [27], from very volatile (network traffic, RAM) to persistent (logs files), they may be heterogeneous in terms of sources (network, system), format, uncertain (incomplete, unclear), not structured (rough data), encrypted. They may also have been falsified. In many cases, they represent a large amount of data to process, i.e. exceed the human processing in a limited time. The main challenge for first responders and analyzers is to assure evidence conditioning and to keep track of all operations they have done.

A. Data collecting and forensic analysis of terminals

One must distinguish tools that can only collect data and those which can also process and analyze them at a first level. There are toolkits from markets that enable to collect digital evidences from the computer (RAM, DISK) while respecting advices for it [28]. The use of market-standardized tools provides generally more guarantees about their reliability and the integrity of data collected (comparing to ad hoc tools developed by experts themselves for ex.). As available tools one can mention the Digital Forensic Framework (DFF) [29], X-ways forensic [30] for live (RAM, registers) and post-mortem (connections, data, metadata, files launched by processes) analysis, Internet Evidence Finder for internet-related data gathering, XRY and UFED Cellebrite for smartphones and GPS terminals [31] [32].

B. System forensic analysis

After intrusion attacks, data have to be collected from network equipment (logs files, traffic) and from the system files. Networks data are generated by tools that have been developed for another usage than security: packets sniffing, traffic analyzing, connectivity testing [33]. In [34] authors have also pointed out the difference between the objectives of auditing tools designers and objectives of forensic analysts. Existing tools have been developed to analyze back tracks (IP addresses, mail counts, web resources) that can be used to give relevant information about the attackers' localization. Other tools enable to analyze files, emails and collect information about systems and running applications in order to prevent spamming [35]. Security Information and Event Management tools are combined tools and platforms designed to collect, analyze, correlate security events in order to produced synthetic reporting [36]. They are event-oriented tools and they use threats databases. They need to be enhanced with data and

knowledge from intrusions tests, attack trees, with knowledge about specific architectures, system configurations and security policies.

Big data architectures associated with virtualization and emulation techniques, data-mining tools such as those based on large graphs constitute a set of processing aids for large amounts of data. Recent tools such as Picviz Inspector [37] are able to process large logs files; they can be viewed as pre-analysis tools, able to reduce the initial entropy of possible ways of analyzing.

C. Attack tree reconstruction

One of the data-gathering interests is to be able to replay the events running by simulation. Events correlation tools can be used to synthetize and reduce the large amounts of IDS-raised alerts and to realize high-level analysis tasks such as foiling attacks plans and scenario, impact analyzing [38]. Some of these tools try to correlate multi-sources indices and events with the aim to go back in attacks and security incidents action-plans [39]. Graphic modeling tools used for the attack trees reconstruction are generally derived from those already used in reliability studies: failure trees, vulnerability trees, attack graphs but they are limited to their static aspect. The dynamic feature of cyber-attacks requires other approaches such as attack modeling with Petri Nets (ex. PENET tool [40]), goal-inducing attack chains [41], which consider events sequences rather than individual events, dynamic Bayesian networks [42] where temporal properties of attacks are considered, adapted attacks trees for systems with dynamic aspects [43] and at last, the Boolean Driven Markov Processes (PDMP) formalism, designed by EDF (Electricité de France) for reliability analysis, which has been proposed for attack trees analysis [44]; for this purpose, the dependency notion (represented by a directed arc in the graph) has been adapted to an attack sequence relation.

D. Synthesis and scientific challenges

The most important issues are: 1) to define a standardized representation language for encoding the events; 2) the intelligent sampling among the large amount of pieces of information (physical pieces, files), sampling that could be expert-driven with his own criteria or semi-automated with specific algorithms (optimization, decision making); 3) information modeling and visualizing in a synthetic way, presenting the analysis outputs in an intuitive mode to help the experts in their reasoning.

V. Forensic Analysis

A. General challenge

Let's consider the general case of a distributed system submitted to an intrusion attack. Most of previously mentioned tools enable to collect indices and tracks, to store and preserve them for processing such as correlations or scenario reconstitution. At this step, the expert has to use his own reasoning to form his own opinion. Only a few works have proposed reasoning tools to help experts, to propose and evaluate hypothesis not only from technical data but also from knowledge about context, behaviours, etc. The scientific challenge is then to build a reasoning scheme that can be able to produce an exploitable report for adjudication, using heterogeneous data that may be uncertain and at different semantic levels. Fig. 3 shows out this process.

The formalism and the tools that are used in causal analysis are generally the same as for diagnosis. But the aim of diagnosis is to identify a faulty component of a system for repairing or replacement. As difference, the forensic process needs to build hypothesis and to verify their plausibility. In [45], the authors propose an approach based on the expert knowledge and that uses fuzzy logic for network forensic. In [46], authors use Bayesian networks to verify hypothesis and constraints in forensic analysis. In [47], authors propose also a Bayesian approach but it is limited to specific attacks.

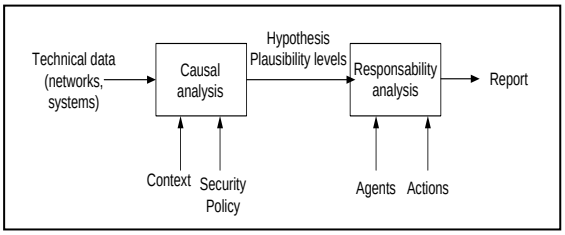


Fig. 3. Causal analysis model

B. Causal Analysis

Causes to effects relations are often represented with causal graphs [48]. Causal Bayesian networks (CBN) provide a suitable and interesting modeling and representation power [49] [50]. The difference with classical BN is that in the case of BN, $A \rightarrow B$ means that the probability of event A, i.e. p(A), has an influence on p(B), which doesn't mean a causal relation. With CBN, it means that A is a cause of B. A probabilistic definition of causality has been defined in [51] but it doesn't integrate any time representation. For cyber-attack, we have previously underlined that the occurrence time of events is an important attribute to use. For that reason other definitions of causality are preferable such as in [52]. In [53], authors make a distinction between endogenous (with random values) and exogenous variables (with fixed values) to establish structural equations of causality. Probabilities are not sufficient to deal with uncertainty. The possibility theory [54] or more particularly possibilistic networks [55] have been proposed to represent and handle uncertainty distribution related to incomplete variables.

C. Responsibility and argumentation logic

Responsibility has not to be confused with causality. If A is a cause and B is an abnormal event, A is the result of an agent action who is the direct cause of the abnormal event or who could have prevented it. Previous works in the artificial intelligence field [56] have proposed logical formalisms to reason about responsibility that emerge from agents behaviours, so that it could be able to answer to questions as: Who is the direct cause of A? What are the most plausible causes of A? Has B a direct effect on A? At what degree? An indirect effect? At what degree is an agent responsible of A?

As there is often a need of explanation for cause and responsibility attributions, an argumentation system is required. In the abstract argumentation approach [57], the argumentation is built using graphs. Nodes represent arguments (which are

elementary objects) and direct arcs represent attack relations. In the argumentation logic approach, the representation is based on logical relations between arguments which have been built from pieces of information [58].

VI. General Synthesis

A. Classes of problems and tools

The data processing approaches required for an efficient detection and investigation have to take into account the characteristics of malicious actions. There are three of them: the willpower of concealment (i.e. to avoid detection), the operating scenario and the individual or collective behaviours that are characteristic of cybercrime classes. For usual forensic challenges, these attributes can be affected as in Table I.

TABLE I Attackers Characteristics

	Concealment	Scenario	Behaviour
Weak signals	X		
Steganography	X		
APT	X	X	
Botnets, rootkits		X	
Fraud		X	
Social Networks	X		X
P2P Networks	X		X

The concealment problem implies to decrease detection levels in such a way to be able to detect weak signals but the induced risk is to increase false positive and false negative ratios. Attacks that use a predefine scenario require the use of a more important amount of data from control of systems and networks, from attacks history, from intrusion tests, and to process real-time correlations. Behaviours aspects need to use data from contexts, sociological studies and expert knowledge.

B. Technological limits

For protection, detection and investigation, the greatest challenge is to develop process chains that can collect and analyze time-limited, sizeable, heterogeneous data about systems, networks and applications. The Table II displays how these characteristics will concern cybercrime cases.

TABLE II Data Characteristics

	Volume	Dynamicity	Heterogeneity
Weak signal	X		X
Steganography	X		
APT	X		X
Botnets, rootkits	X	X	X
Fraud	X	X	
Social networks	X	X	X
P2P networks	X	X	X

The potential information of the available data is not exploited due to technological limits. Data-mining tools (especially classification algorithms) have scalability constraints. The only perspective lies in the big data technology that gives an interesting opportunity to store large volume of data with intelligent query, absorb sporadic input flows without bottleneck effects, and propose adapted analysis and visualization tools, so-called Big Analytics (BA) and Visual Analytics (VA) respectively. To be BA-compatible, algorithms have to be scalable, because they behave linearly vs. data size or because they can be parallelized (some non-supervised clustering processes for ex.) or because they can be adapted to a massive parallelization (scoring methods or neural network-based algorithms).

The visualization methods for multi-dimensional data are generally based on projection operations with the constraint of an efficient interactivity. The forensics needs to correlate variables with time attributes, which requires new approaches based on graphs and graphs matrix [59].

VII. Conclusion

This paper has surveyed the most significant challenges concerning the forensic process as they are presented to the scientific community, especially concerning detection methods and forensic analysis. These challenges are permanently changing with behaviours and action modes. As proposed methods run according to reactive principles they do always lean on a strong survey about cybercrime features. This inventory of methods reveals that the digital forensic process is not addressed as a whole. Supplementary efficiency could probably be gained by designing global responses that involve all required competences.

References

[1] E. Casey, Digital evidence and computer crime: forensic science, Computers and the Internet, Academic Press, 2000

[2] S.J. Wang, C.H. Yang, Gathering digital evidence in response to information security incidents, IEEE Int. Conf. on Intelligence and Security Informatics, Lectures Notes in Computer Science (LNCS), Atlanta, Georgia, USA May 2005

[3] E. Casey, Error, Uncertainty and Loss in Digital Evidence, Int. J. of Digital Evidence, Vol. 1, n°2, 2002

[4] A. Patcha and J-M. Park, An overview of anomaly detection techniques: existing solutions and latest technological trends, Computer Networks, 51, pp. 3448-3470, 2007

[5] IETF, The Intrusion Detection Message Exchange Format (IDMEF), Request for Comments RFC 4765, Internet Engineering Task Force (IETF), 2007

[6] K. Tabia and P. Leray, Bayesian Network-Based Approaches for severe Attack Prediction and Handling IDSs' reliability, Proc. of IPMU'10, pp. 632-642, 2010

[7] H. Tong, Y. Sakurai, T. Eliassi-Rad and Ch. Faloutsos, Fast-Mining of Complex Time-Stamped Events, Int. Association of Computer Investigative Specialist, Forensic procedures, www.iacis.com, accessed in oct. 2012

[8] L. Fillâtre and I. Nikiforov, Asymptotically Uniformly Minimax Detection and Isolation in Network Monitoring, IEEE Trans. On Signal Processing, July 2012, Vol. 60, n°7, pp. 3357-3371

[9] S. Yahi, S. Benharfat and T. Kenaza, Conflicts Handling in Cooperative Intrusion Detection: A descriptive Logic Approach, 22nd IEEE Int. conf. on tools with Artificial Intelligence, ICTAI 2010, Arras, pp. 360-362, 2010

[10] J. Kim, A. Ong and R. Overill, Design of an Artificial Immune System as a Novel Anomaly Detector for Combating Financial Fraud in Retail Sector, Proc. of the Congress on Evolutionary Computation, pp. 405-412, 2003

[11] R.C. Chen, M.L. Chiu, Y.L. Huang and L.T. Chen, Detecting Credit Card Fraud by using Questionnaire-responded transaction model based on support vector machine, Proc. Of the 5th int. Conf. on Intelligent Data Engineering and Automated Learning, Vol. 3177, Oct. 2004, pp. 800-806

[12] M.F.A. Gadi, X. Wang and A.P. Do Lago, Credit Card Fraud Detection with Artificial Immune System, P.J. Bentley, D. Lee, and S. Jung (Eds.): ICARIS 2008, LNCS 5132, pp. 119-131, 2008

[13] S. Panigrahi, A. Kundu, S. Sural and A.K. Majumdar, Card Fraud Detection: A Fusion Approach using Dempster-Shafer Theory and Bayesian Learning, Information Fusion, 2009

[14] R. Cogranne, C. Zitzmann, L. Fillâtre, I. Nikiforo, F. Retraint and Ph. Cornu, Reliable Detection on Hidden Information Based on Non-Linear Local Model, IEEE Workshop On statistical Signal Processing, 4p, 28-30 June, Nice, 2011.

[15] R. Cogranne, C. Zitzmann, F. Retraint, L. Fillâtre, Ph. Cornu and I. Nikiforov, A Cover Image Model for Reliable Steganalysis, 15p, Information Hidding, 18-20 May, Pragua, 2011

[16] M. Latapy, C. Magnien and R. Fournier, Quantifying Paedophile Activity in Large P2P System, Information Processing and Management, 2012

[17] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, Fast unfolding of community hierarchies in large networks, Journal of Statistical Mechanics, 2008

[18] S. Asur, S. Parthasarathy and D. Ucar, An event-based framework for characterizing the evolutionary behavior of interaction graphs, Proc. of the 13th ACM Trans. on the Int. Conf. on Knowledge Discovery and Data Mining (KDD), pp. 913-921, 2007.

[19] F. Gilbert, P. Simonetto, F. Zaidi, F. Jourdan and R. Bourqui, Communities and hierarchical structures in dynamic social networks : Analysis and visualization, Social Network Analysis and Mining, Vol. 1, pp. 83-95, Springer Wien, 2011

[20] M. Oliveira and J. Gama, Understanding Clusters' Evolution, Proc. of Ubiquitous Data Mining (UDM), Workshop in conjunction with the 19th European Conference on Artificial Intelligence - ECAI 2010 in Lisbon, Portugal, August 16-20, pp. 1-6, Lisbon, 2010

[21] T. Aynaud and J.-L. Guillaume, Static community detection algorithms for evolving networks, WiOpt Workshop on Dynamic Networks, pp. 508-514, 2010.

[22] [22] T. Aynaud and J.-L. Guillaume, Multi-Step Community Detection and Hierarchical Time Segmentation in Evolving Networks, Proc. of the 5th SNA-KDD Workshop Social Network Mining and Analysis, August 21, San Diego, 2011.

[23] A. Rosi, M. Mamei, F. Zambonelli, S. Dobson, G. Stevenson and J. Ye, Social sensors and pervasive services : Approaches and perspectives, IEEE Int. Conf. on the Pervasive Computing and Communications (PERCOM) Workshops, Seattle (WA), pp. 525-530, 21-25 March, 2011

[24] K. Yamanishi, J. Takeuchi, G. Williams and P. Milne, On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms, Data Mining and Knowledge Discovery, Vol. 8, pp. 275-300, 2004

[25] National Institute of Standards and Technology (NIST), Guide to Integrating Forensic Techniques into Incident Response, Special publication 800-86, 2006

[26] International Association of Computer Investigative Specialist, Forensic Procedure, www.iacis.com/, accessed in Oct. 2012

[27] IETF, Guidelines for Evidence Collection and Archiving, Request for Comments RFC 3227, Internet Engineering Task Force (IETF), 2002

[28] National Institute of Standards and Technology (NIST), Disk Imaging Tool Specification, V3.1.6, Oct. 2001

[29] Digital Forensic Framework (DFF), http://www.digital-forensic.org/, accessed in July 2013

[30] X-Ways Forensics, http://www.x-ways.net/forensics/, accessed in July 2013

[31] Micro Systemation XRY, http://www.msab.com/, accessed in July 2013

[32] Cellebrite, http://www.cellebrite.com/, accessed in July 2013

[33] N. Meghanathan, S. Reddy Allam and L.A. Moore, Tools and Techniques for Network forensics, Int. J. of Networks Security & Its Applications (IJNSA), Vol. 1, n°1, 2009

[34] S. Peisert, S. Karin, M. Bishop and K. Marzullo, Principles-driven forensic Analysis, Proc. of the New Security Paradigms Workshop, NSPW'05, pp. 85-93, Lake Arrowhead, CA, Sept. 2005

[35] T. Eggendorfer, Methods to identify spammers, Proc. of the 1st Int. Conf. on forensic applications and techniques in telecommunications, Information and Multimedia, e-Forensics'08, 7p, Adelaide, Australia, Jan. 21-23, 2008

[36] AlienVault, http://www.alienvault.com/, accessed in July 2013

[37] Picviz Labs, http://www.picviz.com/, accessed in July 2013

[38] C.Kruegel, F. Valeur and G. Vigna, Intrusion detection and correlation : Challenges and solutions, Springer, 2004

[39] T. Samuel and P.M. Chen, Backtracking intrusions, Proc. of the 9th ACM Symposium on Operating Systems principles, pp. 223-236, NY, USA, 2003

[40] CyberPower, PENET tool, *powercyber.ece.iastate.edu/ penetintro. html, accessed in July 2013*

[41] C.Phillips and L. Painton Swiler, A graph-based system for network-vulnerability analysis, Proc. of New Security Paradigms Workshop, pp. 71-79, 1998

[42] M. Frigault, L. Wang, A. Singhal and S. Jajodia, Measuring Network Security using dynamic bayesian network, Proc. of the ACM workshop on quality protection, QoP'08, Alexandria, USA, pp. 23-30, ACM, NY, 2008

[43] P.A. Khand, System level security modeling using attack trees, Proc. of 2nd Int. Computer Control and Communication Conf. (IC4), pp. 1-6, Feb. 17-18, Karachi, 2009

[44] L. Piètre-Cambacédès and M. Bouissou, The promising potential of the BDMP formalism for security modeling, supplemental volume of the proc. of the 39th annual IEEE.IFIP Int. Conf. on Dependable Systems and Networks (DSN 2009), Estoril, Portugal, June 2009

[45] M.Y.K. Kwan, K.P. Chow, F.Y.W. Law, P.K.Y. Lai, Computer Forensics using Bayesian Network: A case study, HKU CS Technical Report, TR-2007-12, Hong-Kong University, 2007

[46] N. Liao, S. Tian, T. Wang, Network Forensics based based on fuzzy logic and expert system, J. Computer Communications archive, Vol 32, Issue 17, 2009

[47] T. Duval, B. Jouga and L. Roger, XMeta, A Bayesian Approach for computer forensics, ACSAC, Tucson USA, 2004

[48] J. Pearl, Causality: models, reasoning and inference, Cambridge University Press, NY, USA, 2000

[49] A. Darwish, Modeling and Reasoning with Bayesian Networks, Cambridge University Press, NY, 2009

[50] F.V. Jensen and T.D. Nielsen, Bayesian Networks and Decision Graphs, Springer 2007

[51] I.J. Good, A causal Calculus I, British J. of the Philosophy of Science, 11, pp. 305-318, 1961

[52] P. Suppes, A probabilistic Theory of Causality, Amsterdam, 1970

[53] J. Halphern and J. Pearl, Causes and explanations: A structural model approach, part 1: Causes, British J. of the Philosophy of Science 56, pp. 843-887, 2005

[54] D. Dubois, H.T. Nguyen and H. Prade, Possibility Theory, probability and fuzzy sets, in Fundamentals of Fuzzy Sets, D. Dubois and H. Prade Eds, Kluwer Academics Publishers, 2000

[55] S. Benferhat and S. Smaoui, Possibilistic causal networks for handling interventions: a new propagation algorithm, Proc. of the AAAI'07 pp. 373-378, 2007

[56] L. Cholvy, F. Cuppens and C. Saurel, Towards a logical formalization of responsibility, Proc. of the 6th Int. Conf. on AI and Law, pp. 233-242, ACM Press, 1997.

[57] P. Besnard and A. Hunter, Elements of Argumentation, MIT Press, 2008

[58] P.M. Dung, On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games, AI vol. 77, n°2, pp. 321-358, 1995

[59] J. Bertin, Semiology of Graphics: Diagram, Networks, Maps, University of Wisconsin, Press, Madison, (W.J. Berg transl.), 1983

The Phenomenon of Enterprise Systems in Higher Education:Insights From Users

Ahed Abugabah
College of Business Administration
American University
Dubai, United Arab Emirates

Louis Sansogni
School of Business
Griffith University Brisbane,
Australia

Osama Abdulaziz Alfarraj
Computer Science Department
Community College, King Saud
University Riyadh,
Saudi Arabia

***Abstract*—Higher education has been strongly influenced by global trends to adopt new technologies. There has been a call by governments for universities worldwide to improve their performance and efficiency. In response, higher education institutions have turned to Enterprise Resource Planning systems (ERP) in order to cope with the changing environment and overcome the limitations of legacy systems as a means for integration and performance improvement advantages. However, failure rate of ERP implementation is high and debate still exists regarding the various contributions of the ERP systems to performance, especially at the user level, where the core values of information systems are represented and the actual benefits and impacts are created.**

As consequence, this study evaluates the impacts of ERP systems on user performance in higher education institutions with a view to better understand ERP phenomenon in these institutions and determine whether or not these systems work well in such a complex environment. The study also developed and validated statistically a new model suggesting a more inclusive sight for examining ERPs utilization and impacts and combining the key ideas of three well-known and widely used information systems models.

Keywords—ERP systems; user performance; system quality; higher education

I. Introduction

In order to deal with today's changing environment and overcome the limitations of legacy systems, organizations have turned to Enterprise Resource Planning (ERP) systems [1; 2]. ERP systems have been adopted by numerous organizations and support most industries including airlines, telecommunications and education. These systems fundamentally represent the most significant development in terms of costs and corporate use of information systems [1; 3].

ERP systems permit the seamless flow of information across the entire organization and address the problem of fragmentations of information or "Islands of information" in organizations [1]. Since the emergence of ERP packages in the late 1990s, they have become popular among practitioners and IS researchers alike [2]. According to Gartner Inc., the revenue for ERP software around the world in 2011 is US $253.7 billion; this amount represents an increase of 7.5% compared to 2010's, as stated in [3].

Although ERP systems are being used widely around the world with billions of dollars and countless hours spent in implementing such systems, they bring along many problems and weaknesses. Most of these implementations are unsuccessful and fail through "inadequate adoption", this being just one of a number of failure factors [4]. As such, debate still exists regarding the various contributions of the ERP systems to performance, as the failure rate of implementation is high. In particular, failure to respond to the users' needs leads many large information systems' projects, including ERP ones, to fail making it impossible to realize the expected benefits in today's rapidly evolving business environment [5; 6].

Consequently, especially in the last decade, ERP systems have begun to attract the attention of researchers. Little attention however, has been given to their adoption in higher education. Organizations, including higher education, aim at obtaining value from ERP systems through improved overall efficiency, operational efficiency and business efficacy [7; 8].However, this possibility remains unclear in higher education, foreshadowing the need for investigations to see whether or not ERP systems deliver sustainable outcomes and whether or not they help to improve performance and enhance productivity.

II. ERP Systems In Higher Education

Higher education institutions have been strongly influenced by global trends to adopt new technologies. There has been a call by governments for universities worldwide to improve their performance and efficiency [4]. Challenges including increasing expectations of stakeholders such as students and governments, decreasing governmental support, meeting quality and performance requirements, and maintaining competitive education environments have pressured universities to adopt new strategies in order to improve their performance [5]. In response to the many challenges faced, higher education institutions have turned to ERP systems in order to cope with the changing environment [6] and to replace aging management and administration computer based systems [7]. This consequently would improve learning services by providing better managerial tools [8] thereby increasing their pace of organizational change and effectiveness [5; 9].

ERP systems in universities can provide academic entities including schools and departments with completely functional applications for research and teaching [10].

They improve information access for planning and managing the institution, and enable users to access students' information, academic records and other data needed to complete their daily work [11], leading thus to improved business processes and services provided to the faculty, students and employees [8; 12].

Even though, implementation of ERP systems in higher education institutions is often described as extremely difficult. Expenses and risks involved are high. It is also sometimes unsuccessful or ineffective, whereas the return on investments is medium to long-term. Research on ERP systems in higher education reported a large number of failures and/or inadequate adoption of ERPs [13]. For example, EDUCAUSE conducted series of studies to assess ERP systems for tertiary institutions [8; 14] reporting King that 50% of these implementations were over budget and over timeline schedules. Recent research claimed that as many as 60% to 80% percent of all ERP systems fail due to lack of meeting expected outcomes [15] and/or lack of performance improvement, with users expressing dissatisfaction with their performance.

Although ERP systems are the largest information systems' project adopted by universities, with significant resources allocated to implementation(e.g. higher education institutions spent more than five billion dollars in the last few years on ERP investments) [11], little research has been conducted on ERP implementations in universities compared to other environments [9]. At best, these studies brought about the identification of a number of critical factors related to the ERP implementation in higher education such as, staff training [10; 4], leadership and culture [15], change management, system functionality [16], ERP integration with education processes [14; 17; 18; 19; 20], the evolution of ERP systems and the university curriculum [21; 11], and the ability of ERP systems to support business processes in universities [22].

Overall, prior research on ERPs revealed some important results, opening the path for new empirical investigations on ERP systems in a university environment, especially using ERP systems in the classroom for learning and teaching purposes [23] such as the usefulness of using ERP systems for enhancing learning by providing ways to transform the classroom into a real business environment [24].

With the spotlight of prior research mainly focused on success and failure factors and the successfulness of the ERP implementation other important issues at the user level such as user evaluation and user performance of ERPs thus remain elusive [25]. As consequence, this study focuses on ERPs' users to evaluate the impacts of ERP systems on user performance in higher education institutions with a view to better understand ERP phenomenon in these institutions. The study was designed to answer several questions related to how ERP systems affect user's performance, whether ERP systems improve performance, and identify the most significant factors that affect user performance within the context of ERP systems in higher education.

III. Research Approach and Methodology

A survey methodology was used to gather data from ERP users in universities. The questionnaire was synthesized after an extensive review of the IS and ERP literature. The study was carried out in 6 large universities in Australia implemented several modules of ERP systems in different functional units such as human resource, students administration and finance. The respondents numbered 387 ERP users in total from various functional areas in these universities. The name of the university is withheld due to our non-disclosure agreement with the executives.

Measurement items used in the operationalization of the instrument were adopted from relevant prior research as listed in Appendix 2 [26; 27; 28]. The factors investigated in this study consolidated the main factors investigated in three well known IS models as illustrated in the study model below

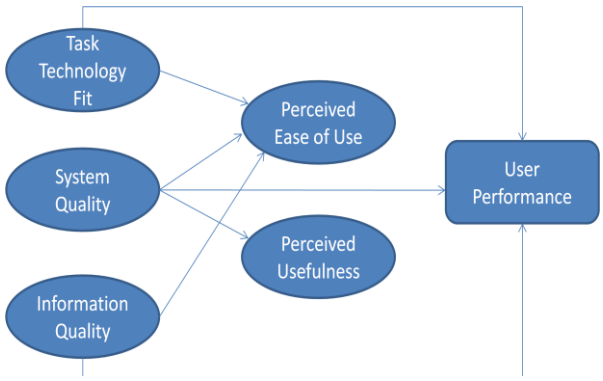


Fig. 1. The study model

A. Validity and Reliability

The reliability and validity of the measurement instrument was carried out using reliability and factor analysis. The entire instrument, as well as the individual variables, achieved high levels of reliability. The results showed that the reliabilities of the constructs (Cronbach's coefficient α) ranged from 0.84 for PU to 0.97 for UP, indicating high reliability, as mentioned in Table I (Appendix 1).

Instrument validity: Factor analysis was carried out to examine measurement construct validity. Typically construct validity is considered to be satisfactory when items load high on their respective constructs (factors). The cut-off point used in this analysis was .5, as recommended by [30; 31]. All correlations below this point were considered low. As shown in Table I (Appendix 1), all items had high loadings on their respective factors, with most of items above 0.70, demonstrating high construct validity.

IV. Discussion and Findings

Multiple regression analysis (MRA) was used to assess the effects of various factors included in this study on user performance. MRA is useful to identify significant contributions made by these factors to predicting user performance, rather than correlations, by themselves, as they do not explain all the relationships between factors in a study.

This is considered to be appropriate given that most prior studies similar to this study have used regression analyses to assess the relationships between the factors and thus doing the

same here will facilitate comparison of the study results with previous ones [32; 33; 28].

To evaluate the overall impact of ERP systems on user performance, a multiple regression analysis was performed in two stages. The first stage utilized TTF, SQ, IQ, PEOU and PU as independent factors, while the second stage used only the PEOU and PU factors. It should be remembered, however, that user performance (UP) was self-assessed, so the UP measure reflects self-user performance.

The results of the analyses are presented in Table II(Appendix 1) and indicated that the whole model has a significant positive relationship with user performance. As shown in Table II(Appendix 1) ($F = 20$), which is significant at the ($P < .01$) level. This relationship is strong ($R = .782$), explaining 61.2% of the variance in user performance and thus hypothesis H1 is supported, which is consistent with the results of previous ERP systems' research (such as [34]).

Regression weights, especially β values, can be used to compare the individual contributions of independent factors [35]. The results showed that IQ did not contribute significantly and uniquely to explaining user performance beyond the explanation provided by the other independent factors, while TTF contributed significantly to user performance. The largest unique and significant contribution for predicting user performance was provided by SQ ($\beta = .604$). An overall multiple regression was also conducted, using all factors, including PU and PEOU, to check the contribution of each factor in explaining user performance in the presence of PU and PEOU. In other words, the second test was to explain how PU and PEOU mediated the systems' impact on user performance, as shown in Table II (Appendix 1).

A regression analysis was also employed to identify the factors that contribute significantly to predicting user performance. The importance of each factor was also displayed in Table II (Appendix 1). The significant contribution for each factor is shown in column labelled B. PU was the strongest protector ($\beta = .425$), explaining 71.7% of the variance in user performance, while the second factor was SQ ($\beta = .407$), explaining 58.5% of the variance in user performance. Thus, PEOU plays a critical role in increasing the impact of ERP systems on user performance.

The findings demonstrated that ERP systems impact user performance in higher education significantly and positively. The results of the analysis showed that all above mentioned factors contribute to user performance and explain a high percentage of the variance in user performance. However, this result was further explored to identify the contribution for each factor on the variance in user performance.

Previous studies, which investigated ERP impacts on users indicated that SQ and IQ are very important factors that affect benefits of use [36]. They however mentioned that SQ plays more important roles than its IQ counterpart in terms of influencing ERP benefits. In this sense, [37] found that SQ and IQ are considered the most important factors when evaluating ERP systems' impacts. Others mentioned that SQ and system integration are two important factors that contribute towards the formation of the overall systems' impacts. This study demonstrated the importance of all of the abovementioned factors and explored the relative contribution of each to user performance.

ERP users reported that their performance is improved through experience thereby improving efficiency and effectiveness. The results showed that users believe that ERP systems provide high quality information, which helps reduce errors and solve performance problems when they occur. Furthermore, SQ and TTF play an important role in enhancing performance quality and increasing the quantity of work performed by users. The results showed a satisfactory level of fit between ERP systems and users' needs, and tasks requirements, considering the sophisticated characteristics of ERP systems. It was found that users believe that ERP systems have the required compatibility and learn ability that helps them in performing their tasks.

Considering the unique characteristics of ERP systems, the current study contributes greatly in the identification of the determinants of user performance in practice. That is, the design of ERP systems should fit exactly the tasks that users are engaged with. The findings of the study suggest that TTF can affect PEOU and PU directly. For example, the more flexible and convenient ERP systems are, the more they will be perceived as useful PU and easy to use PEOU.

In other words, ERP users believe that ERP systems are helpful in supporting the performance of their tasks and also easy to use if they are designed to be applicative and fit to their tasks. Careful consideration of users' needs and task requirements in a specific industry will help guide system designers and practitioners with the design and implementation of ERP systems, in light of the diversity of vendors, designers, ERP systems functionalities, and industries [38].

V. Conclusion

This study investigated the impact of enterprise resource planning systems on user performance in higher education. The findings of this study, for the most part, are consistent with previous studies on ERP systems such as [34; 32; 28], and several others that have extended TAM and TTF models, though there are also some differences between our results and the others such as [32]. The results of the study demonstrated that PU has a positive direct and indirect impact on system use and user performance. In fact, among all factors investigated, PU has a large effect on and/or mediated the impacts of other factors on performance. This result is significant as it shows that in a complex IS environment, just as in non-complex environments, PU of a system is perhaps more important than its ease of use [32]. Thus, designing attempts focused on enhancing PU of the ERP systems will be worthwhile since it is more likely to lead to more system impacts and an improved performance. Unlike previous studies that used PU measures to investigate the system impacts in most cases and/or user satisfaction to measure system success and system usage [40; 41], the current study provides significant progress in measuring systems' impacts and user performance.

The implications of the results revealed that meaning beyond the information obtained from the ERPs, the fitness and consistency between ERP applications and work aspects,

the exact meaning of information obtained from the ERP, and the correct meaning of the information on the ERP systems were all important factors in predicting user performance. Participants of the study indicated that these factors were adequate to handle the work processing needs and led consequently to improved performance. For example, adequacy of the ERP system has a significant effect on user performance. Therefore, the design of an ERP system's interface and functionality must be aligned with user needs and task requirements and also should be easy to navigate among different ERP modules.

This study provides more clarifications and explanations about the potential benefits and outputs of ERP systems for users. This is becoming important as organizations, especially with the increasingly huge investments on ERP systems' installations made by higher education institutions. The little empirical research in this environment available especially at user level, associated with lack of empirical research in this environment made the benefits of the systems unrecognized yet [42; 43]. The study is deemed to be useful in explaining how users can obtain values from ERP systems and reflect them in their task and job accomplishments.

REFERENCES

[1] Eric, W., Ngai, Chuck, C., & Law, H. (2007). An investigation of the relationships between organizational factors, business process improvement, and ERP success *An International Journal, 14*(3), 287-406.

[2] Davenport, T. (1998). Putting The Enterprise into The Enterprise Systems. Harvard Business Review, 76(4), 121-132.

[3] Ifinedo, P. (2011). Examining the influences of external expertise and in-house computer/IT knowledge on ERP system success. *The Journal of Systems and Software 84*, 2065-2078.

[4] Allen, D., & Kern, T. (2001). *Enterprise Resource Planning Implementation: Stories of Power, Politics, and Resistance* at the Proceedings of the IFIP TC8/WG8.2 Working Conference on Realigning Research and Practice in Information Systems Development: The Social and Organizational Perspective Idaho.USA.

[5] Fisher, M. D. (2006). Staff Perceptions of an Enterprise Resource Planning System Implementation: A Case Study of three Australian Universities. Unpublished PhD Thesis, Central Queensland University, Queensland.

[6] McCredie, J., &Updegrove, D. (1999). Enterprise System Implementations: Lessons from the Trenches. *CAUSE/EFFECT, 22*(4), 1-10.

[7] Pollock, N., &Cornford, J. (2001). Customising Industry Standard Computer Systems for Universities: ERP Systems and the University as an 'Unique'OrganizationEngland: UMIST.

[8] Kvavik, R., Katz, R., Beecher, K., CARUSO, J., & KING, P. (2002). The Promise and Performance of Enterprise Systems for Higher Education. *EDUCAUSE, 4*, 5-123.

[9] Nielsen, J. (2002). Critical success factors for implementing an ERP system in a university environment: A case study from the Australian HES. Griffith University, Brisbane.

[10] Watson, E., & Schneider, H. (1999). Using ERP in education *Communications of AIS, 1*(9), 12-24.

[11] Davis, M., &Huang, Z. (2007). ERP in Higher Education: A Case Study of SAP and Campus Management. *Issues in Information Systems, VIII* (1), 120-126.

[12] King, P., Kvavik, R., & John, V. (2002). Enterprise Resource Planning Systems in Higher Education. *EDUCAUSE, 22*, 1-5.

[13] Botta-Genoulaz, V., & Millet, P. (2006). An investigation into the use of ERP systems in the service sector. *International Journal of Production Economics, 99*(1), 202-221.

[14] Judith, P. (2005). Good Enough! IT Investment and Business Process Performance in Higher Education ECAR, key findings, 2005

[15] Mehlinger, L. (2006). *Indicators of Successful Enterprise Technology Implementations in Higher Education* Unpublished Doctorate Thesis Morgan state University, Morgan state.

[16] Vevaina, P. (2007). Factors affecting the implementation of enterprise systems within government organizations in New Zealand. Auckland University of Technology, Auckland

[17] Cynthia, L., & Harold, W. W. (Writer) (2004). Appropriating Enterprise Resource Planning Systems in Colleges of Business: Extending Adaptive Structuration Theory for Testability.

[18] Todd, J., Alden, C. L., James, M., & Jon, O. (Writer) (2004). A Customized ERP/SAP Model for Business Curriculum Integration.

[19] Casper, D., & Dirk-Jan, S. (Writer) (2004). Best Practices of Business Simulation with SAP R/3.

[20] Jarmoszko, A. T., & Michael, G. (2004). Choosing an ERP-type System for a Belarus Enterprise. *Journal of Information Systems Education, 15*(3), 255.

[21] Paul, H., Brendan, M., & Andrew, S. (Writer) (2004). Second Wave ERP Education.

[22] Jane, F., Ulric, J. G., Jr., Catherine, U., & George, H. (Writer) (2004). Twelve Tips for Successfully Integrating Enterprise Systems across the Curriculum.

[23] Yvonne Lederer, A., Gail, C., Glenn, S., & Albert, L. H. (Writer) (2004). Enterprise Systems Education: Where Are We? Where Are We Going?

[24] Noguera, H. J., & Watson, F. E. (Writer) (2004). Effectiveness of using an enterprise system to teach process-centered concepts in business education.

[25] Lope Ahmad, R., Othman, Z., &Mukhtar, M. (2011). Campus ERP implementation framework for private institution of higher learning environment in Malaysia. *WSEAS Transactions on advances in engineering education 1*(8), 1-12.

[26] Goodhue, D., & Thompson, R. (1995). Task-technology fit and individual performance. *MIS Quarterly, 19*(2), 213-233.

[27] Delone, W., & McLean, E. (1992). Information systems success: the quest for the dependent variable. *Information systems research 3*(1), 60-95.

[28] Calisir, F., &Calisir, F. (2004). The relation of interface usability characteristics, perceived usefulness and perceived ease of use to end - user satisfaction with enterprise resource planning systems. *Computer in Human Behavior 20*(505-515).

[29] Golafshani, N. (2003). Understanding Reliability and Validity in Qualitative Research *The Qualitative Report, 8*(4), 597-607

[30] Goodhue, D., Klein, B., & March, S. (2000). User evaluations of IS as surrogates for objective performance. *Information & Management, 38*, 87-101.

[31] Hong, K., & Kim, Y. (2002). The critical success factors for ERP implementation: an organizational fit perspective. *Information & Management, 40*(25-40).

[32] Amoako-Gyampah, K. (2007). Perceived usefulness, user involvement and behavioral intention: an empirical study of ERP implementation. *Computers in Human Behavior, 23*(3), 1232-1248.

[33] Lucas, H., &Spitler, V. (1999). Technology use and performance: A field study of broker workstations. *Decision Sciences, 30*(2), pp 291-322.

[34] Kositanurit, B., Ngwenyama, O., &Osei-Bryson, K. (2006). An exploration of factors that impact individual performance in an ERP environment: an analysis using multiple analytical techniques. *European Journal of Information Systems, 15*, 556-568.

[35] Pallant, J. (2007). *SPSS Survival Manual: A step by step guide to data analysis using SPSS* (3rd Ed.). New South Wales: Allen &Unwin.

[36] Chien, S., &Tsaur, S. (2007). Investigating the success of ERP systems: Case studies in three Taiwanese high-tech industries. *Computers in Industry, 58*(8), 783-793.

[37] Ifinedo, P., &Nahar, N. (2006). Quality Impact and Success of ERP Systems: A Study Involving Some Firms in the Nordic-Baltic Region. *Journal of Information Technology Impact 6*(1), 19-46.

[38] Yen, D. C., Wu, C.-S., Cheng, F.-F., & Huang, Y.-W. (2010). Determinants of users' intention to adopt wireless technology: An empirical study by integrating TTF with TAM. *Computers in Human Behavior, In Press, Corrected Proof.*

[39] Wu, J.-H., Chen, Y.-C., & Lin, H.-H. (2004). Developing a set of management needs for IS managers: a study of necessary managerial activities and skills. *Information & Management, 41*(4), 413-429.

[40] Wu, J., & Wang, W. (2006). Measuring KMS success: A re-specification of the DeLone and McLean's model. *Information & Management, 43*(6), 728-739.

[41] Kwahk, K.-Y., &Ahn, H. (2009). Moderating effects of localization differences on ERP use: A socio-technical systems perspective. *Computers in Human Behavior, In Press, Corrected Proof*.

[42] Sun, Y., Bhattacherjee, A., & Ma, Q. (2009). Extending technology usage to work settings: The role of perceived work compatibility in ERP implementation. *Information & Management, 46*, 351-356.

[43] Hellens, L., Nielsen, S., &Beekhuyzen, J. (2005). *Qualitative case studies on implementation of enterprise wide systems*. Hershey: Idea Group Publishing.

[44] Goodhue, D., & Thompson, R. (1995). Task-technology fit and individual performance. *MIS Quarterly, 19*(2), 213-233.

[45] DeLone, W., & McLean, E. (2003). The DeLone McLean model of information system success: a ten-year update. *Journal of Management Information Systems, 19*(4), 3-9.

[46] Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Management Information System Quarterly 13*(Sep), 318-340.

[47] Davis, F., Bagozzi, P., &Warshaw, R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Management Science 35*, 982-1003.

[48] Goodhue, D., Klein, B., & March, S. (2000). User evaluations of IS as surrogates for objective performance. *Information & Management, 38*, 87-101.

APPENDIX 1

TABLE I. RESULTS OF FACTOR ANALYSIS AND MEASUREMENT RELIABILITY (N = 387) *

Factors/Items	Loading	Mean	SD	Factors/Items	Loading	Mean	SD
TTF(α=.90)		4.9	.96	Corr1	.75	3.2	.93
Loc1	.74	5.3	1.39	Corr2	.60	3.3	.92
Loc2	.81	4.9	1.39	**PU (α=.84)**		3.9	.78
Com2	.74	5.4	1.12	PU1	.69	3.9	.83
Com3	.75	5.3	1.14	PU2	.67	4.2	1.03
ITsub2	.84	4.7	1.33	PU3	.76	3.7	.92
ITsub3	.85	4.8	1.31	PU4	.73	3.7	.98
Ade1	.84	4.8	1.34	**PEOU(α=.89)**		3.3	.89
Ade2	.60	4.8	1.36	PEOU1	.72	3.2	1.00
Mea1	.74	4.5	1.30	PEOU2	.85	3.2	.97
Mea2	.78	4.3	1.30	PEOU3	.89	3.4	.98
IQ (α=.87)		3.6	.61				
Accees1	.71	3.5	.90	**UP(α=.97)**		4.5	1.14
Access2	.82	3.4	.91	Effici1	.81	4.6	1.28
Complet1	.50	3.4	.88	Effici2	.77	4.9	1.34
Complet2	.50	3.7	.76	Effici3	.76	4.7	1.29
Tim1	.53	3.6	.86	Effici4	.76	4.6	1.32
Tim2	.69	3.6	.87	Effici5	.65	4.6	1.24
SQ (α=.87)		3.3	.63	Effici6	.78	4.7	1.32
Integ1	.77	3.1	.85	Effici7	.74	4.8	1.35
Integ2	.78	3.3	.83	Effici8	.69	4.7	1.34
Integ3	.58	3.2	.99	Effec1	.715	4.5	1.38
Relia1	.66	3.7	.87	Effec2	.61	4.4	1.32
Relia2	.83	3.6	.79	Effec3	.60	4.7	1.30
Restime1	.73	3.3	.96	Crea1	.91	3.9	1.52
Restime2	.74	3.2	.94	Crea1	.83	3.7	1.57

*Only loadings of 0.5 or above are shown; ** Values in parenthesis represent Cronbach's alpha

TABLE II. ERP IMPACTS ON PERCEIVED USER PERFORMANCE

Independent Factors *	*User performance*					*User performance through PEOU & PU*				
	B	**S.E**	***β***	**t**	**Sig**	**B**	**S.E**	**β**	**t**	**Sig**
TTF	.198	.056	.167	3.54	.005	.09	.048	.07	1.41	.06
IQ	.146	.093	.077	1.56	.110	.14	.080	-.07	-.25	.07
SQ	1.079	.083	.604	13.0	.001	.73	.075	.407	9.228	.001
PU						.62	.052	.425	11.315	.001
PEOU						.19	.049	.149	3.620	.001
R		.780					.850			
R^2		.610					.730			
F		201					210			

*TTF: Task technology fit, IQ: information quality, SQ: system quality, PU: perceived usefulness, PEOU: perceived ease of use, UP: user performance.

TABLE III. THE SIGNIFICANT CONTRIBUTION OF EACH FACTOR IN PREDICTING USER PERFORMANCE

Factors*	***R***	***R^2***	***Adjusted R^2***	***β***	***S. E***
SQ	.765	.586	.585	.407	.73578
PU	.847	.718	.717	.425	.60771
PEOU	.857	.731	.728	.153	.59500

APPENDIX 2

Constructs	Measurement items	Source
Task technology fit **		[44], [26]
Locatability	1. It is easy to determine what application is available and where to do my job.	
	2. It is ease to locate the data in the ERP applications that I use.	
Compatibility	1. ERP applications that I use are consistent with my tasks.	
	2. ERP applications fit with my work aspects.	
Meaning	1. The exact meaning of information obtained from the ERP, relating to my task, is easy to find out.	
	2. The correct meaning of the information is obvious and clear on the ERP software	
Adequacy	1. The ERP software that the university has meets my task requirements.	
	2. The ERP software is adequate to handle my work processing needs.	
IT support	1. I get the kind of quality computer-related training that I need.	
	2. The IT people I deal with understand my work objectives.	
	2. It is easy to get IT support and advice from IT people when I use ERP applications.	
Information quality *		[27], [45]
Accuracy	1. Our ERP system provides me with accurate information.	
Relevancy	1. Our ERP system provides relevant information.	
Timeliness	1. Our ERP system provides me with the information I need in a timely manner.	
	2. The information in our ERP system is timely and regularly updated.	
	3. Getting information from our ERP system on time improves my work quality.	
Completeness	1. I can find complete information when I need it in our ERP system.	
	2. The information in our ERP system is sufficient to do my work.	
Accessibility	1. The information in our ERP system is easily accessible.	
	2. Information in our ERP system is easy retrievable.	
	3. Convenience of information in our ERP system saves my time in my job.	
Perceived usefulness *		[46], [47]
	1. Our ERP system is useful for my job performance.	
	2. I cannot accomplish my job without the ERP system.	
	3. Our ERP system supports me in attaining my overall performance goals.	
	4. Our ERP system makes it easier to do my job.	
Perceived ease of use*		[46], [47]
	1. Our ERP system is user friendly.	
	2. It is easy to learn how to use our ERP system.	
	3. I find the ERP system is easy to use.	
System quality *		[44], [48]
Reliability	1. Our ERP system is reliable.	
	2. Our ERP system has consistent information.	

Correctness	1. I find it easy to correct the errors related to my work by using our ERP system.	
	2. Our ERP system helps me reduce the errors in my job.	
Response time	1. Our ERP system reacts and responds quickly when I enter the data.	
	2. Our ERP system responds quickly to my inquiries.	
Integration	1. Our ERP system allows for integration with other systems.	
	2. Our ERP system effectively combines data from different areas of the university.	
	3. Our ERP system is designed for all levels of user.	
User performance**		[27], [44],
Efficiency	1. I can accomplish my work quickly because of the ERP system quality.	[32]
	2. Our ERP system lets me do more work than was previously possible.	
	3. Our ERP system has a positive impact on my productivity.	
	4. Our ERP system reduces the time taken to accomplish my tasks.	
	5. Our ERP system increases the cases I perform in my job.	
	6. Using our ERP system in my job enables me to accomplish tasks more quickly.	
	7. Overall, our ERP system improves my efficiency in my job.	
	8. Our ERP improves my performance quality.	
Effectiveness	1. Our ERP helps me solve my job problems.	
	2. Our ERP reduces performance errors in my job.	
	2. Our ERP system enhances my effectiveness in my job.	
Creativity	1. Our ERP helps me to create new ideas in my job.	
	2. Our ERP system enhances my creativity.	
	3. Overall our ERP system helps me achieve my job goals.	

Loop Modeling Forward and Feedback Analysis in Cerebral Arteriovenous Malformation

Y.Kiran Kumar[1a]*,
[1]Philips HealthCare,
[a] Research Scholar, Manipal University

Shashi.B.Mehta[2]
[2]Philips IP&S, Philips Innovation campus, Bangalore, India

Manjunath Ramachandra[3]
[3]Philips Research, Philips Innovation campus, Bangalore, India

***Abstract*—Cerebral Arteriovenous Malformation (CAVM) hemodynamic in disease condition results changes in the flow and pressure level in blood vessels. Cerebral Arteriovenous Malformation (CAVM) is an abnormal shunting of vessels between arteries and veins. It is one of the common Brain disorder. In general, the blood flows of cerebral region are from arteries to veins through capillary bed. This paper is focus on the creation of a new electrical model for spiral loop structures that will simulate the pressure at various locations of the CAVM Complex blood vessels. The proposed model helps Doctors to take diagnostic and treatment planning for treatment by non-invasive measurement.. This can cause rupture or decreased blood supply to the tissue through capillary causing infarct. Measuring flow and pressure without intervention along the vessel is big challenge due to loop structures of feedback and forward flows in Arteriovenous Malformation patients. In this paper, we proposed a lumped model for the spiral loop in CAVM Structures that will help doctors to find the pressure and velocity measurements non-invasively.**

Keywords*—*Vessel Loops; AVM; Lumped Model

I. Introduction

Cerebral Arteriovenous Malformation (CAVM) is an abnormal tangle of brain blood vessels where arteries shunt directly into veins with no intervening capillary bed which causes high pressure and he morrhage risk. Intracranial Arteriovenous malformations (AVM) constitute usually congenital vascular anomalies of the brain. AVMs are composed of complex connections between the arteries and veins that lack an intervening capillary bed. A brain modeling of the Hemodynamics with physical properties of Cerebral AVM is important in understanding the dynamics of pressure flow relationships and implications of alterations in these properties with respect to, pressure monitoring, and logical approach to therapy and treatment.

The aim of this work is to model the pressure at various vessel loops analysis of a MRA/3DRA dataset using the Lumped models. The input parameters used for the proposed simulation and for analysis is the clinical parameters [1-3]. In the present work we have used new modeling approach for the Vessels Loops from 2D & 3D data and also proposed the modeling for the forward loop and feedback loop in any vessel structure .In the literature analysis, the modeling is based on the fluid dynamics of the vessel. It has some drawbacks on the analysis using various signals. In the lumped modeling, the analysis is based on the electrical circuit analogy using WindKessel models, was used provide a computationally simple way to obtain information about the overall behavior of the Neurovascular system. The authors had proposed electrical parameters and derived a number of lumped models for blood flow and pressure variances in the Cerebral Arteries Windkessel as well as lumped parameter models are used to simulate pressure and blood flow in the arterial system. In these models, electric potential and current are analogous to the average pressure and flow rate, respectively. A particular vessel (or group of vessels) is described by means of its impedance, which is represented by an appropriate combination of resistors, capacitors and inductors. The vessel loop analysis and modeling is a gap, where very few authors have analyzed on this research due to high complexity behavior of vessels [4-7]. . The below figure 1.0, shows the complexity of CAVM, that indicates the complex anatomy of the Vessel.

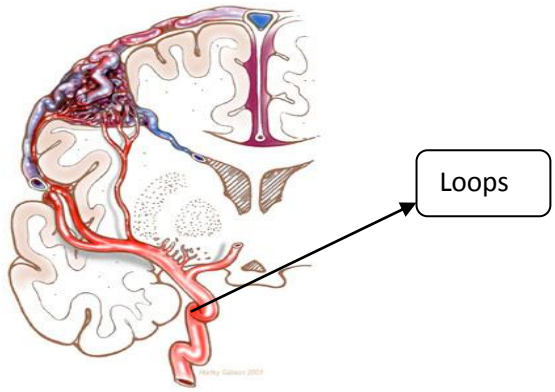


Fig. 1. Cerebral Arteriovenous Malformation (CAVM)

The problem statement we are targeting in the proposed work is to address the above gap, by using the WindKessel model for the vessel Loops types like forward and feedback loop modeling using lumped model for the asymmetrical and symmetrical networks that simulates the actual neurovascular complexity.

II. Methodology

The proposed methodology is based on the 3DRA dataset, which is obtained from Philips Allure Unit. The input 3DRA image is used as the input volume of the Brain AVM (BAVM). The following steps are applied to volume image:

1) Acquisition of 3DRA/MRA dataset of AVM.

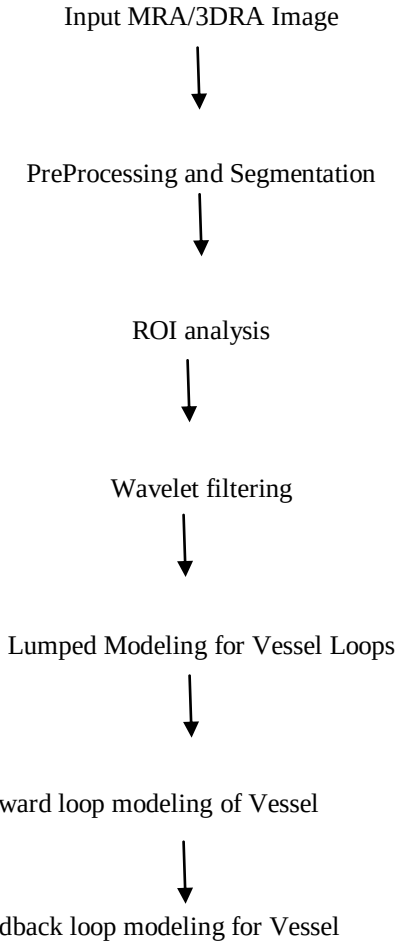
2) Segmentation and Preprocessing techniques are applied to the dataset and enhanced contrast, smoothing algorithm and edge detection algorithm based on intensity.

3) The filtering is applied to remove the noise using wavelet transform.

4) 3D ROI is drawn for the Vessel loop region, which automatically propagated to all the slices, by applying interpolation technique.

5) The modeling is constructed for the loop region for the segmented 3D-ROI. Depending on the clinical scenario of the patient, the loop is separated in to feedback /forward loop of blood flow.

The flow chart



A. Lumped Model Analysis:

Windkessel and lumped models are often used to represent blood flow and pressure in the arterial system. These lumped models can be derived from electrical circuit analogies where current represents arterial blood flow and voltage represents arterial pressure. Resistances represent arterial and peripheral resistance that occur as a result of viscous dissipation inside the vessels, capacitors represent volume compliance of the vessels that allows them to store large amounts of blood, and inductors represent inertia of the blood. The windkessel model was originally put forward by Stephen Hales in 1733 and further developed by Otto Frank in 1899 [8-10]. , the equivalent RLC values are calculated using the following equations:

$$R = \frac{8l\pi\mu}{A^2} \quad (1)$$

Where μ is blood viscosity, l and A are in respect length and cross section area of each arterial segment. Blood viscosity is a measure of the resistance of blood to flow, which is being deformed by either shear stress or extensional stress. This simulation has considered because blood viscosity will cause resistance against Blood flow crossing.

$$L = \frac{9l\rho}{4A} \quad (2)$$

Where ρ is blood density.

$$C = \frac{3l\pi r^3}{2Eh} \quad (3)$$

Where r, E, h are in respect artery radius, Elasticity module and thickness of arteries. The arterial radius and thickness are calculated from the segmented vessel, which is used for calculating the R, L, C values to generate electrical Model [8-10].

B. Creation of Loop Model:

The below figure 2.0 shows the complex loop of the vessels of the AVM patient. The loop is segmented using threshold based and filtered output is shown in figure 3.0

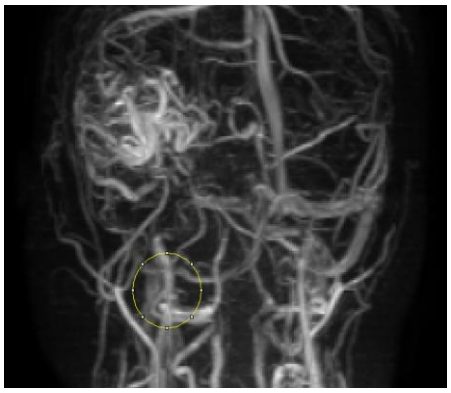


Fig. 2. MRA Dataset

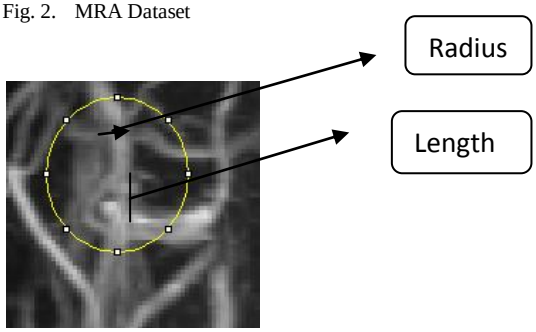


Fig. 3. Preprocessed and Segmented loop

The challenge in the modeling is that it is not direct spiral loop structure, present in the Cerebral vessel, it is the complex structures, which is modeled for the various forward and feedback loop is created using RLC networks, the circuit shows the loop modeling with variation in the inductors and capacitance as shown in figure 4.0 [11-13]:

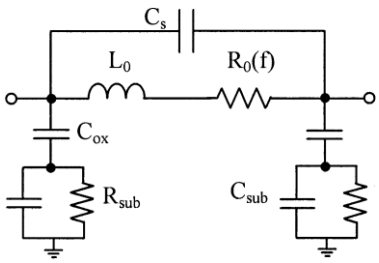


Fig. 4. Modeling Circuit

III. METHODOLOGY

The Loop Modeling is analyzed using various inputs and simulated for various input pressure range /voltage sources. The table 1.0 shows the variations across various inputs.

TABLE I. VARIATIONS ACROSS INPUTS

Input Voltage/ Pressure mmHg/Volts	Output Pressure mmHg/Volts	Clinical results /Application results mmHg/Volts	Mechanical Simulation mmHg/Volts
0.8Volts	0.78	0.8	0.78
1.2 Volts	**1.1**	**1.2**	**1.15**

The statistical analysis is performed for various input voltage range to verify for the stability of the model. The correlation coefficient shows that the results are correlating with mechanical and with actual results from Clinical sites.

TABLE II. STATISTICAL ANALYSIS

Sample size	17
Correlation coefficient r	0.9896
Significance level	P<0.0001
95% Confidence interval for r	0.9705 to 0.9963

IV. RESULTS AND DISCUSSION

The implementation is done using MATLAB Software using 3D-RA and MRA dataset. The electrical model is derived for the vessel loops of forward and feedback propagation of the blood flow based on the symmetric and asymmetrical networks using clinical parameters. The clinical parameters are analyzed and converted in to electrical parameters which help to create a WindKessel model using R, L, C values using Electrical Analogy conversion. The RLC values and its combinations are modified based on the type of network and also type of vessel loops.

The figure 5.0 shows the Matlab implementation of the modeling of loops.

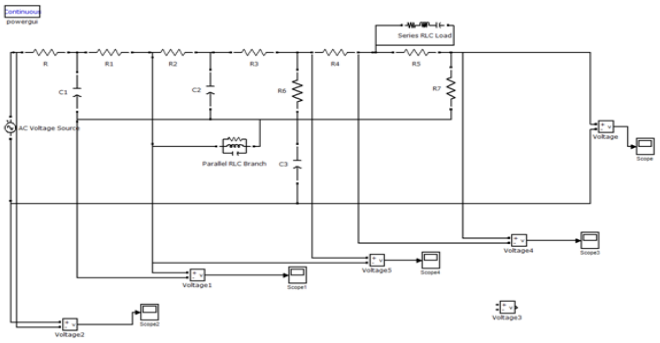


Fig. 5. MATLAB Implementation

This model is simulated and validated with clinical results and with mechanical outputs. The mechanical simulation is performed using ANSYS software. The figure 5.0 shows the MATLAB implementation to simulate the vessel loop structures. The effect of vessel loops analysis shows the clinical significance on the pressure and flow rate variation of the blood flow in AVM patients.Simulations were also performed using the phantoms and also using 2D –DSA images that are varying length with networks of symmetric and asymmetrical . In these simulations for various values of diameters are used to simulate the actual clinical scenario.

V. CONCLUSION

In this paper, we have proposed a neurovascular vessel loop modeling which indicates the blood vessel movement and other clinical parameters variation. This work is based on Lumped Model and it is validated through the Mechanical Model. The future scope of this work is to analyze all the types of vessel loops for complete brain.The modeling outputs are used by Doctors for the diagnosis, treatment planning for the AVM Surgery and also for the AVM management. This work is in progress for various organs and complex vessels loop analysis.

REFERENCES

[1] Persuasive technology: using computers to change what we think and do, Fogg, B.J, Morgan Kaufmann Publishers, Boston, 2003, 30-35

[2] Review of Zero-D and 1-D Models of Blood Flow in the Cardiovascular System Yubing Shi, Patricia Lawford, and Rodney Hose, Biomed Eng Online. 2011; 10: 33. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3103466/

[3] Can Induction of Systemic Hypotension Help Prevent Nidus Rupture Complicating Arteriovenous Malformation Embolization?: Analysis of Underlying Mechanisms Achieved Using a Theoretical Model, Tarik F. Massoud, George J. Hademenos, William L. Young, Erzhen Gao, and John Pile-Spellman, AJNR Am J Neuroradiol 21:1255–1267, August 2000.

[4] Modelling cerebral haemodynamics: a move towards predictive surgery,Edlong ,March 9, 2007,Thesis Report.

[5] Theoretical modelling of arteriovenous malformation rupture risk: a feasibility and validation study,.Erzhen Gao a, William L. Young,a Department of Electrical Engineering, Columbia University, New York, NY 10027, USA,b Department of Anesthesiology, College of Physicians and Surgeons of Columbia University, New York, NY 10032, USA, IPEM, 1998.

[6] Can Induction of Systemic Hypotension Help Prevent Nidus Rupture Complicating Arteriovenous Malformation Embolization?: Analysis of Underlying Mechanisms Achieved Using a Theoretical Model, Tarik F. Massoud, George J. Hademenos, AJNR Am J Neuroradiol 21:1255–1267, August 2000.

[7] Experimental model of intracranial avm in the acute stage , Shinichi, Department of Neurology, Medical university, Fukushima, Neurological Medical Chir (Tokyo), 288-293, 2005.

[8] AVM compartments, Do they modulate Nidal Pressures? An electrical network analysis, Litao, AJNS, 173-180, December 2012.

[9] Frequency-Independent Equivalent-Circuit Model for On-Chip Spiral Inductors, Yu Cao, IEEE Journal of Solid-State Circuits, Vol. 38, NO. 3, March 2003.

[10] On deriving lumped models for blood flow and pressure in the systemic arteries, Olufsen MS, Nadim A.MBE, Department of Mathematics, North Carolina State University, 2004,61-80.

[11] A lumped parameter model of coronary flow to analyze time intensity curves extracted from angiograms Electrode-rail dielectrophoretic assembly effect: formation of single curvilinear particle-chains on spiral microelectrodes, Xiaolu Zhu, Microfluidics and Nanofluidics October 2010, Volume 9, Issue 4-5, pp 981-988.

[12] A Simple Electrical Equivalence Model of Intracranial Cerebrospinal Fluid Pulsatility: Design and Validation in Healthy Normals., Jieun, Published in Proc. MIUA 2006.

[13] Frequency-Independent Equivalent-Circuit Model for On-Chip Spiral Inductors , Yu Cao, IEEE Journal of Solid –State Circuits Vol 38, NO. 3, March2003

Design and Evaluation of Spatial Multi Interaction Interface

Chang Ok Yun
Arcade Game Regional Innovation Center
Dongseo University
Busan, Korea

Tae Soo Yun, YoSeph Choi
Division of Visual Contents
Dongseo University
Busan, Korea

***Abstract*—Nowadays interactive displays are capable of offering a great variety of interactions to users thanks to advancement of ubiquitous computing technologies. Although many methods of interactions have been researched, usability of the devices is still limited and they are offered only to a single user at a time. This paper proposes a spatial multi-interaction interface that can provide various interactions to many users in an ambient environment. An interaction surface is created for users to interact through IR-LEDs Array Bar. The coordinate information of the hand is extracted by detecting the area of the hand of a user in the interaction surface. Then users can experience various interactions through "spatial touches" on the interaction surface. In our paper a usability evaluation is carried out for our new interface which gives the emphasis to the interaction interface. The usability of our new interface is shown to be significantly better through statistical testing using t-testing. Finally, users can perform various interactions with natural hand motions only without the aid of devices that have to be operated manually.**

Keywords—Interactive display; Ambient environment; Interaction surface; Spatial interaction

I. INTRODUCTION

Until very recently, development of ubiquitous computing technologies has been focusing on building infrastructures such as constructing communication networks in the areas where human activities are possible, but now attention is focused on multi-modal interaction technologies with which humans can interact naturally with the computing environment built by the communications infrastructure. To construct an "ambient" or "disappearing" computing environment" which characterizes the ubiquitous display, a system has been developed which offers different interactions depending on the distance between users(Stephanidis, 2009). The existing technology only offers either a simple touch type of display or grasps the user's intention using various sensors, but now "Gossip Wall" makes use of distance sensors, and the interactive ambient display system collects and shares the position information of the user through various cameras. The problem with the interaction devices that use the existing technology is that it is unable to provide natural interactions to users. Moreover, it is a sensor recognition method; therefore, users have to install sensor recognition devices. Thus many users cannot participate all together. In other words, since the existing technology uses traditional devices such as keyboards and mouse, most interactions are limited to certain areas that are directly related to these devices. To overcome these limitations of the existing techniques, it is now necessary to develop a new technology that can control the components more naturally and intuitively in an ambient environment as well as a new method of showing relevant information more effectively to users.

This paper proposes an interactive system with which humans can interact using simple hand movements alone without the aid of sensor recognition devices in a ubiquitous ambient environment. The proposed system differs from the existing interaction and the operation of space interaction. Whereas in the traditional technology, a single user interacts using an interaction device (a mouse) that is connected to a computer, in the proposed spatial multi-interaction system, a number of users can interact in the space with a variety of contents by using only simple hand movements. This paper is organized as follows. Section 2 describes previous researches on various interaction techniques in an ambient environment. Section 3 describes the design and implementation of the proposed interface. Section 4 describes the result of the proposed technique and the experimental environment. Section 5 gives a comparative analysis of the proposed system and the existing system. Finally, Section 6 closes the paper with conclusions.

II. RELATE WORKS

The Ambient Intelligence (AmI) provides electronic environments that are sensitive and responsive to the presence of people. The ISTAG played a decisive role in the further development of the AmI vision and launched a scenario planning exercise to demonstrate what might be realised through AmI technology (IST Advisory Group, 2003). The users are provided with applications and services with which they interact in unobtrusive manner (Aarts, Harwig & Schuurmans 2003). In the AmI, the sensors are used to be supported and utilized in the interaction, and applications and services should be consistent, easy to handle and easy to learn. Furthermore, devices are wirelessly connected and form intelligent networks which create environments in which people are surrounded by intelligent and intuitive interfaces that are embedded in all kinds of objects (Holmlid, S. & Björklind, A. (2003)). And the AmI will be performed with the aim of hiding the presence of technology to the users, providing seamless and unobtrusive interaction paradigms.(Stephanidis, 2009).

The University of Toronto developed an interactive ambient display for public use, which interacts with users in four stages in accordance with the distance between the user and the

display. Users can have a different interaction in each stage (Vogel & Balakrishnan, 2004). However, interactions are limited because the user must wear many types of recognition devices, and it is difficult for multiple users to use the system. Norbert A. Streitz, et al. of Fraunhofer IPSI has attempted to advance interactions from human-computer interaction (ICI) to human-environment interaction (HEI). Through their research products such as Roomware(Tandler, 2004) and Ambient Agoras(Streitz, 2007), they tried to integrate the space made up of actual structure which has inherent computing ability and virtual information space. Stanford University's iRoom (Johanson, 2002) has presented a new type of interacting method whose purpose is to develop a display for public use in cooperative environment. Gossip Wall developed by Fraunhofer IPSI in Germany is one of the many techniques of using various devices that have been researched. When the user is in an ambient environment that is very far away from the display, general information is displayed, and when the user comes close to the display and the display recognizes the user, the display shows contents that are relevant to the user. When the user comes very close and walks into the interaction area, the user can use the display of the mobile device to get detailed information. It is also possible to connect with outside. In the case of Gossip Wall, the user uses a mobile device that employs a distance sensor system to get the needed information, and the display grasps the position of the user. However, the problem with Gossip Wall is that it can trace the user only when the user carries the mobile device. Another problem is that the available display devices are limited in variety.

Recently in Korea, various researches are in progress on the application of infrared LED to the sensing systems. Tracking technique that makes use of infrared LED is applied to the mutual interactions between the interface and users (Kim & Kim, 2009), and a rehabilitation training system is being developed by applying the infrared LED band (Park & Park, 2008). In the case of Nintendo Company, the immersive user interface (Yoon et al., 2009) is offered by tracing the gaze and the location of the user, applying the Wii controller and infrared LED. In another case, a variety of contents are offered through a simple pointing and controlling device (Hong et al., 2009; Baek et al., 2005; Park & Park, 2008) that uses infrared LED. In still other case, FTIR (Frustrated Total Internal Reflection)-based tabletop displays and interactive wall displays (Choi et al., 2008) are developed, and they are being applied to various contents. In short, the existing interaction interface using the infrared LED contains inconvenient and restricting elements.

Touchless interaction techniques (Barrée et al., 2009), by allowing user to employ hand gestures, remove the burden related to physical contact and promote natural interaction with digital information made tangible through large display surfaces. Touchless interaction can also be multimodal: in this case the interaction events embrace different human senses (visual, auditory and olfactory). Most of the emergent game devices come from the entertainment industry, such as: Nintendo's Wii Remote Controller, Microsoft's Project Natal and Sony Play Station3 Motion Sensing controller.

There are more intuitive motion-based interactions such as detection of natural body movement. Besides providing a more intuitive game playing (Vaughan-Nichols, 2009) (not only based on button pushing), these devices also stimulate the development of touchless interfaces that go beyond interaction styles using WIMP (Windows, Icons, Mouse and Pointers) elements. Don't Touch Me(Bellucci et al., 2010) is a system providing the users with the possibility to collaborate, generate and place multimodal annotation on a digital map using Nintendo's Wiimote. And the rapid prototyping of touchless-enabled interfaces(Lee, 2008) is providing the feasibility of the Nintendo's Wiimote. A 3D vision-based ambient user interface(Hong & Woo, 2006) as an interaction metaphor that exploits a user's personal space and its dynamic gestures is the system of touching the augmented SpaceSensor. The eye-gaze input system(Murata, 2006) is using as one of the touchless interaction interface. The system is providing conditions such as the moving distance, size of a target, and direction of movement in a pointing task.

Consequently, the user either has to carry simple pointing or control devices or touch the display directly. This paper proposes a spatial multi interaction interface in which users can interact with the display through the interaction surface created at the IR-LEDs Array Bar using only natural hand motions without any devices that have to be operated manually..

III. SPATIAL MULTI INTERACTION INTERFACE

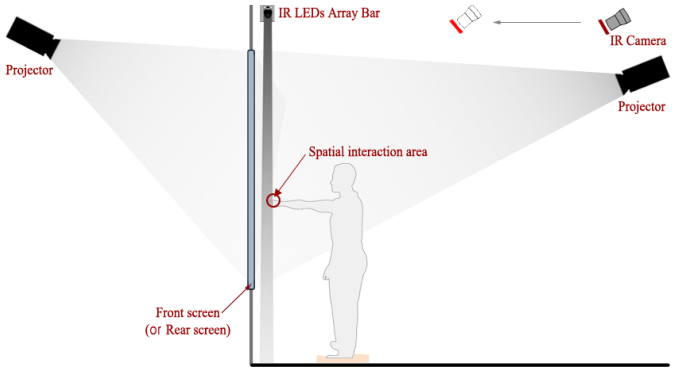


Fig. 1. The overall system layout.

Fig. 1 shows the overall layout of a large scale interactive display which offers a variety of spatial multi interactions to users. This system consists of a large projector-based wall-surface display device, IR LEDs Array Bar that uses an IR-LEDs generating device, and an IR Camera equipped with an IR-Filter. As can be seen in Fig. 1, the user interacts by using hand motions in a natural way on the IR interaction surface, but not on the restricted screen of the existing display. In the environment of our system, the distance from the user and screen is changed according to IR LEDs Array Bar location. Spatial interaction area in the range to reach the hands (the range of people reaching out), the interaction is enough. In addition, our system arranged the camera and projector in the front side of user considering the case where the person competes for the projector and the screen used the Rear Screen and solved the phenomenon that this system is covered by the user body.

And this system was shrouded through the location change of the camera (Adjust the distance between the camera and screen, the camera focal length adjustment), it concluded the problem.

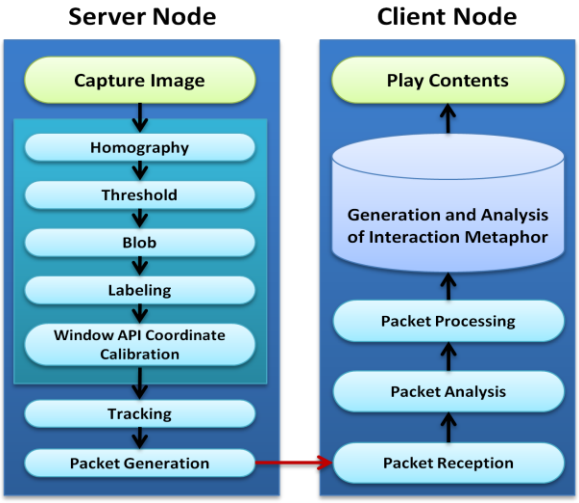


Fig. 2. Flowchart of the System.

Fig. 2 shows the overall flowchart of the proposed spatial multi-interaction interface. First, the IR camera inputs the image. Since the image obtained through the camera is generally distorted, the image is revised by using the homography matrix. To remove any noise other than a certain intensity of illumination, the threshold process is used to remove noise. After the coordinates are revised in the blob-labeling process, the position of the hand is located through tacking. This information on the hand coordinates applies spatial multi-interaction through the client (contents) and network communications.

A. *Environmental Setup for Image Getting*

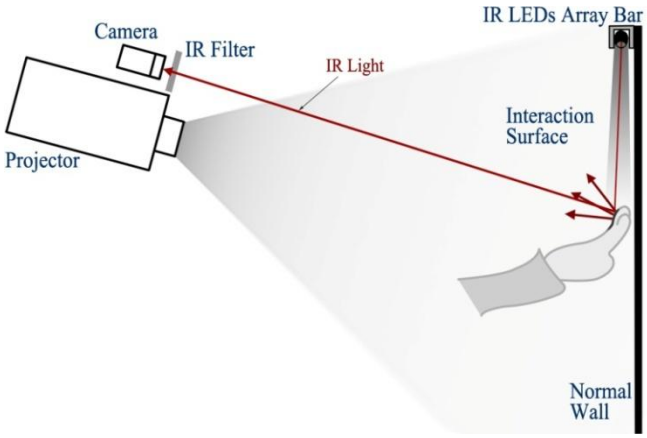
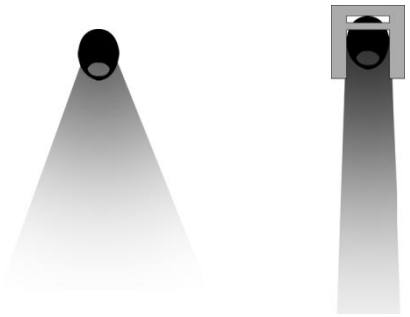


Fig. 3. Positions of the Projector and Camera.

The proposed system does not require a special screen, but an IR-LEDs array bar is installed on the upper end of the screen in proportion to the width of the screen. The distance of the projector is determined in proportion to the size of the screen. The bigger the screen, the further away is the projector from the screen, and the smaller the screen, the closer to the screen is the projector positioned. Thus, it can be installed any place using various spaces. Moreover, since the proposed system uses the touchless mode, it can recognize the user's gestures in space and makes it possible for the user to interact in a natural and comfortable way. In other words, even when the interaction surface is separated from the wall, the user can do spatial interactions on the interaction surface and enjoy a variety of contents. The camera is placed above the projector so that it can see the projected area on the screen. At this point, the images coming out of the projector are cut off by using the Band-pass filter because without the use of Band-pass filter, all images projected by the projector would appear on the screen. The Band-pass filter cuts off areas smaller than 850nm [see Fig. 3].

In this work, the IR-LEDs array bar was installed to fit the width of the screen for spatial multi-interaction interface. The IR-LEDs array bar was constructed by arranging IR LED with intervals of 3cm. The screen and infrared rays must be kept horizontal so that images of everything and anything unnecessary except the realm of the user's hand are removed. To maintain the horizon, the radiation scope of lights was narrowed by cutting off the vicinities of IR LED. To prevent infrared rays from radiating to 45°, both edges of LED were cut off so that rays would radiate vertically. Fig. 4(b) shows an IR-LEDs array bar using light straight. To track the user's hand, the IR-LEDs array bar radiates light rays vertically downward from the top of the screen. When the light rays reach the user's hand, light rays cannot travel in straight lines because of the hand. As a result the light rays that lost the ability to travel in straight lines remain on the back of the user's hand, and the infrared camera obtains the location of the hand. The infrared LED used for this purpose radiates light waves of 45° 880mm.



(a)ordinary IR LED (b) IR LED Array Bar

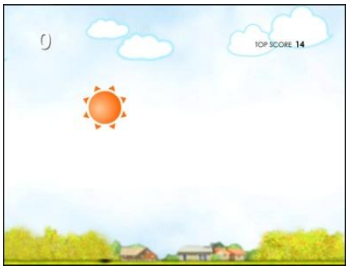
Fig. 4. IR LED Array Bar that uses the principle of light rays traveling in straight lines.

B. *Image Distortion Correction*

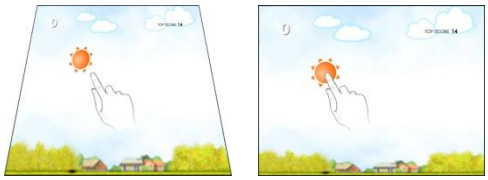
In general, a system that uses a projector and a camera requires a stage of revision to get the coordinate value of the exact position.

If the coordinate for the hand is calculated on the basis of the distorted image as shown in Fig. 5(b), applying it to the screen coordinate, it is not possible to get the right result. Therefore, after getting the image input by the camera, the keystones for the image to be projected through the projector are corrected and the homographic relations are calculated for

the images to be projected. Following these procedures, it is possible to obtain information on the exact coordinate of the image projected by the camera as shown in Fig. 5 (c). There are two ways of calculating conversion relations between the camera's image and the projector's image: The fundamental matrix that calculates the epipolar geometry between the two images projected by the camera and the projector, and the homographic matrix that measures the relations of dots between two planes. The fundamental matrix deals not with the relations between dots but with the relations between dots and planes. Thus it suffers from severe noise because it has to consider many planes, and consequently, it produces more noise than the homographic matrix. Taking these problems into consideration, this paper applies homography to calculate the conversion relations between the camera and the projector. With these findings, distorted images are corrected.



(a) Input image by the projector



(b)Distorted input image (c) Input image corrected

Fig. 5. Input image corrected by homography.

The homographic method is capable of calculating geometric projective relations between two planes as follows. The relation between the input image's coordinate $X_i' = (x_i', y_i', w_i')^T$ and the converted output image can be presented with the following equation.

$$\begin{pmatrix} x_i' \\ y_i' \\ w_i' \end{pmatrix} = \begin{pmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{pmatrix} \begin{pmatrix} x_i \\ y_i \\ w_i \end{pmatrix} \quad (1)$$

Eq. (1) can be simplified as $X_i' = HX_i$ where H is a homography with a size of 3×3, and Eq. (1) can be transformed as Eq. (2).

$$X_i' H X_i = 0 \quad (2)$$

Eq. (2) is developed with $Ah = 0$, and A can be calculated as follows:

$$A = \begin{bmatrix} 0 & 0 & 0 & -x_1 & -y_1 & -1 & y_1'x_1 & y_1'y_1 & y_1' \\ x_1 & y_1 & 1 & 0 & 0 & 0 & -x_1'x_1 & -x_1'y_1 & -x_1' \\ 0 & 0 & 0 & -x_2 & -y_2 & -1 & y_2'x_2 & y_2'y_2 & y_2' \\ x_2 & y_2 & 1 & 0 & 0 & 0 & -x_2'x_2 & -x_2'y_2 & -x_2' \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & -x_n & -y_n & -1 & y_n'x_n & y_n'y_n & y_n' \\ x_n & y_n & 1 & 0 & 0 & 0 & -x_n'x_n & -x_n'y_n & -x_n' \end{bmatrix} \quad (3)$$

$$h = (h_1, h_2, \cdots, h_8, h_9)^T \quad (4)$$

h is 9×1 vector which is H matrix arranged in dictionary sort, and A is the $n\times9(n\geq4)$ matrix which is the combination of X_i and X_i'. A has 8 degrees of freedom and so it requires at least 4 pairs of coordinates. h consists of Eigenvectors which correspond to the smallest Eigenvalue of A^TA. The elements of h thus calculated are substituted with each element of homography in order. Fig. 6 shows the homographic relations of the corresponding points between the images of the screen, the projector, and the images.

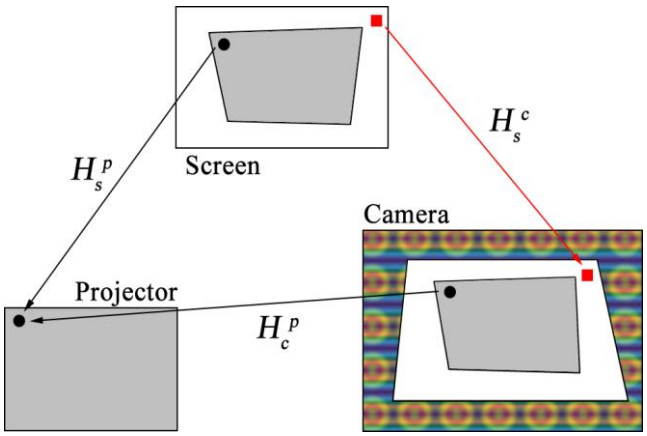


Fig. 6. Projector-Camera Homography Relations.

H_s^p, the homographic relations between the images to be projected by the camera and the images to be projected by the projector, must be calculated in order to correct the keystones of the images to be projected after getting the input from the camera. If the boundary line ratios that form the 4 vertexes on the screen and the coordinates of the vertexes that are fixed on the camera are known, the homography H_s^c can be calculated.

In addition, if the rates are known for the images that are to be projected from the projector and if the coordinates of the four vertexes that are fixed on the camera are already known, H_c^p, the homography between the camera and the projector can be calculated. Therefore, H_s^p, the homography between the screen and the projector can be obtained via Eq. (5) if H_s^c, the homography between the screen and the camera, H_c^p, the homography between the camera and the projector H_c^p are already known.

$$H_s^p = H_s^c H_c^p \tag{5}$$

C. Correction of Mouse Coordinate

Coordinates obtained through homography calculation are the coordinates in the realm of the camera's view, and they are not the coordinates that are actually applied to the window's coordinate. If the camera's entire view area is 640*480 and the resolution is 1024*768, the error range will inevitably be big. To reduce the margin of errors in the position information, the mouse's range is calculated at the API level and the window's coordinate is controlled.

$$X_k = X_i \times \frac{65535}{width}$$
$$Y_k = Y_i \times \frac{65535}{height} \tag{6}$$

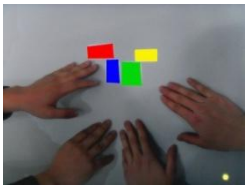
Since the actual coordinate of the mouse has the area of 0 ~ 65535 (up, down, right and left) regardless of the resolution, the current resolution 65535 is divied by the width and height and multiplied by the coordinate (X_i, Y_i) which is the result of homography calculation. Then the right coordinate is calculated. (Refer to Eq. (6))

D. Multi-interaction for publicly shared space

A spatial multi-touch function is essential for various interactions. Spatial multi-touch means that the input information appears simultaneously in many parts of the screen. In the proposed system, the information that is input through the band pass filter is binary coded [Fig. 7(a)] and the noise is removed. Each position is located in the blob-labeling process [Fig. 7(c)].

Computers being used currently do not have the device to simultaneously input many kinds of information and they are incapable of performing a spatial multi-interaction function. For this purpose, the network communication is used to communicate with the system and the contents. The coordinate sought out by the infrared camera forms packets and transmits the contents in real time. The contents analyze the packets transmitted and implement the spatial multi-interaction function.

(a) Recognition of the finger contact area



(b) Bob treatment of the detected area

(c) Implementation of the multi-interaction function applied to the applied contents

Fig. 7. The hand area recognition and Blot treatment of relevant area using infrared rays and the result of Blob treatment of the relevant area.

IV. Experimental Interface and Results

A. Experimental Set-Up

The proposed system was implemented using Microsoft Visual C++, OpenCV on a system in which a Pentium IV 1.8GHz, 2GB RAM is installed. After obtaining the camera's images, treating the images, correcting the distorted images, and recognizing coordinates by using Visual C++ language, it is now possible for users to interact with the contents through the network communication. In the proposed system, the camera gets the information on the user's position by recognizing the IR rays reflected on the interaction surface that is created by the IR-LEDs Array Bar. The surrounding environmental conditions for the proposed system can influence the performance of the system; therefore, the experimental environment was designed by eliminating these negative environmental factors. Osram 880nm IR LED was used to construct the IR-LEDs Array Bar. An 850nm IR filter was used as a means of inputting the IR images only. To get the IR area images effectively inputted, the IR cutoff filter attached to the front of the CCD was removed before the experiment. All the images coming out of the projector are cut off by using a band pass filter (the filter that cuts off waves shorter than 850nm), and only the IR rays are accepted. As a result, when the user puts out a hand toward the interaction surface in order to interact the contents, the infrared light coming down from the IR-LEDs Array Bar are cut off because it cannot penetrate the hand. At this point, the camera finds out the coordinate information by processing the infrared light reflected from the hand into images. The IR-LEDs Array Bar was constructed by arranging infrared LED with 3cm intervals. To enforce straight traveling, both sides of infrared LED were cut off so that the infrared light that used to radiate out to 45° will only travel vertically downward. Fig. 8 presents the overall experimental environment for the proposed system.

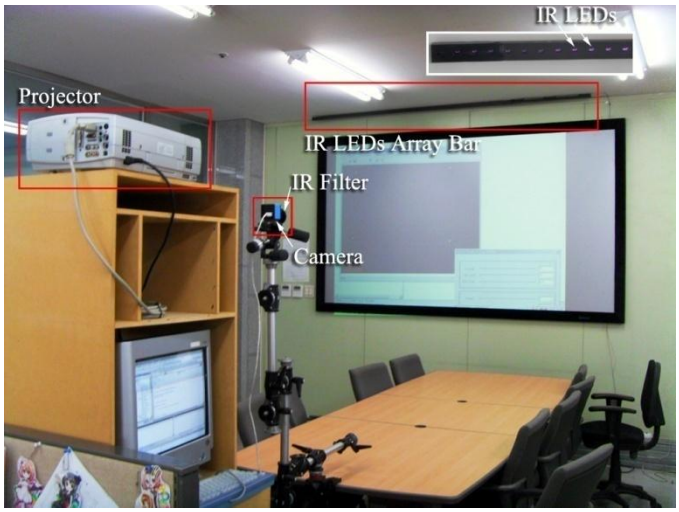


Fig. 8. The experimental environment for the proposed system.

B. Results of Performance Evaluation

The functions of devices composing the recognition system are important factors determining the function of the entire interaction face because the recognition function of the interaction interface influences the usability.

The existing system that typically can be easily accessed by the user is pressure-sensitive touch screen. The differences between the proposed and existing systems are as follows:

TABLE I. DIFFERENCES BETWEEN PROPOSED AND EXISTING SYSTEM

Item	***Proposed system***	***Existing system***
Recognition method	Camera based	Pressure-sensitive
Size	More than 120 inches	17, 24, 42 inch
Multi Touch	Possible	Possible(one dual touch)
Strength	High scalability(Large) Wall type Lower cost Easy to implement	Convenient access Easy to use
Weakness	Need to learn about the basic functions	Not scalable

In this study we calculated the FPS (frames per second) of touches made by the user to measure the function of the entire recognition system. Hands on comparing the number of existing systems can be up to two touching a proposed system to support two or more the number of touches. In other words, the speed can be faster for two, but that does not support more than the number of touch.

TABLE II. TOUCH RECOGNITION SPEED(FPS)

Number of hands		1	2	3	4	5	6	7	8	9	10	11	12
Recognition speed (fps)	***Proposed system***	30	30	30	30	30	30	29	29	29	29	29	29
	Existing system	50	50	-	-	-	-	-	-	-	-	-	-

The high intensity of the light from the IR LED BAR is so strong that it creates a film. Light that is emitted at a certain wavelength and intensity spreads evenly over the interaction film. Thus interactions were achieved on the entire interaction area. The result of the camera frame test showed an efficiency of 30fps over the entire area, and it was confirmed that the interactions were tracking naturally without pause to the eyes of the user. Furthermore, as shown in Fig. 9, many users can interact in space on a large scale interactive display environment. A high efficiency of 21-30 fps was maintained while multi-users were interacting.

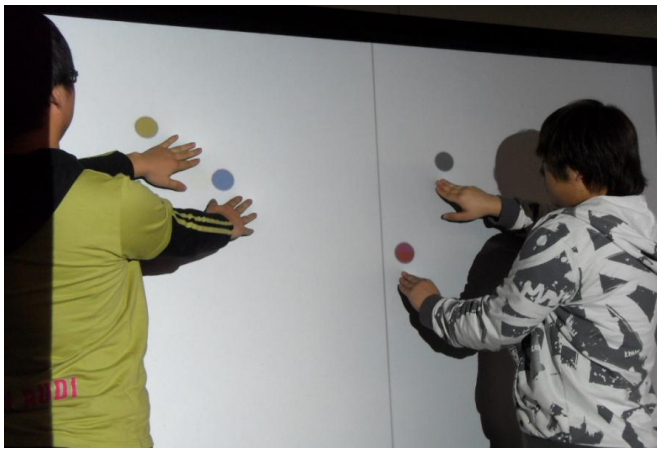


Fig. 9. Spatial multi-interaction demonstration.

To evaluate the recognition ability, the speed of hand area recognition was measured. It takes approximately 0.01msec to capture an image through the recognition camera, including the time required for basic image capturing, pre-treatment of every recognition process, and GUI generating process. As listed in Table 2, the time required for recognizing the hand motion is about 0.01msec for as many as 9 hand areas. When the hand areas increase to 10, the time required to recognize the image increases to 0.02msec.

TABLE III. HAND AREA RECOGNITION TIME

Number of hands		1	2	3	4	5	6	7	8	9	10	11	12
Recognition time (msec)	***Proposed system***	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02
	Existing system	0.005	0.005	-	-	-	-	-	-	-	-	-	-

V. EVALUATION OF USABILITY

The proposed system offers an interface with which the user can interact in space on an ambient environment display. This system is expected to be fabricated and implemented more easily and naturally than the existing multi-touch method. In the proposed system, interaction is made possible by the position of the hand whereas in the existing multi-touch screen system, interaction is carried out by touching the screen directly. To examine how the spatial multi-interaction interface is received by the users, this section evaluates the usability compatibility of the game contents developed for the proposed system with those of the existing multi-touch screen method. Thus, the user's satisfaction is measured.

A. Purpose of Evaluation

According to the research purpose of the newly developed ambient environment display, we compared the proposed system with the existing touch display (type 1) and measured the degree of satisfaction the users felt on "ease of use" and "learnability" of the two displays.

1) Ease of Use : "How easily and freely can users use the contents?"is the question to answer here. Ease and freedom include both the convenience of inputting viewed from the perspective of hardware and the user's sensory perception of control in using the contents. Contents of the game are implemented totally by the user's choice and control. The inconvenience or limitations felt by the users give the users an impression that they are not in complete control of the game, and perceived inconvenience and limitations immediately lower the user's interest in the contents. The 'Ten Usability Heuristics' of Jakob Neilsen widely applied to the evaluation of software and websites also point out that usability is on the top of the list of requirements for interface.

2) Learnability of Use : "High learnability"of contents means that the user can use the contents intuitively with the least of amount of learning. Many options in using the contents and non-intuitive methods require users to learn many things, and this is directly linked to lowering of users' interest. Previous studies including Jacob Neilsen's 'Recognition rather than recall'(Neilsen, 1994) all emphasize the importance of usability. Especially in the games characterized by play and immediate interaction, an increase in the amount of learning "how to play?" plays the role of a fatal factor that lowers the interest in the contents.

B. Method of Evaluation

Ease of use and learnability of the ways to play the contents that apply the proposed system are compared to those of the existing multi-touch screen method to measure the degree of satisfaction felt by the users. The 50 developers evaluated using the Heuristic Evaluation method prepared in advance. Developers who participated in the evaluation were composed of professionals(almost workers) and amateurs(almost students). Through questionnaires to the participants before the experiment existing touch screen interface and a development interface for use in the survey was conducted, the user enough information on how to use them after the experiment was performed as described. For each interface using the time to consider the content Running time was 5 minutes. The amount of playing the contents is totally three times.

TABLE IV. EVALUATION PARTICIPANTS(50 DEVELOPERS

Item	*Online game developers*	*Console game developers*	*UI developers*	*Network developers*	*Web contents developers*
Count	16	12	8	8	6

Heuristic Evaluation method (Neilsen, 1994) is more efficient in cost, evaluation time, and manpower required for evaluation than other evaluation methods because exact heuristics are applied on the objects to be measured. Since the purpose of this evaluation is to find out the strengths as well as the weaknesses of the game contents that are newly developed, the Lickertis scale was used to record the degree of satisfaction which the participants felt on all the items on the list. In this evaluation, the scale of 1-7 is applied instead of the 1-5 scale which is more widely used, since in the 1-5 scale, results tend to center around 3 ("fair" satisfaction, neither good nor bad).

The game content used for the evaluation is a matching game for two or more nursery school age children in which a dam is rebuilt and destroyed. The party who scores more points within a given time is the winner. The higher scoring determines the winner and the loser. The game is played by multi-touch inputting and dragging.

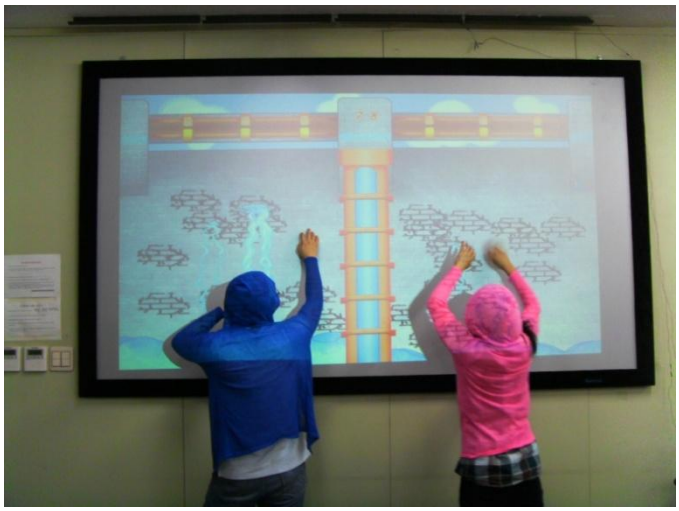


Fig. 10. Experimenting multi-touch matching game.

C. Establishing the evaluation heuristics

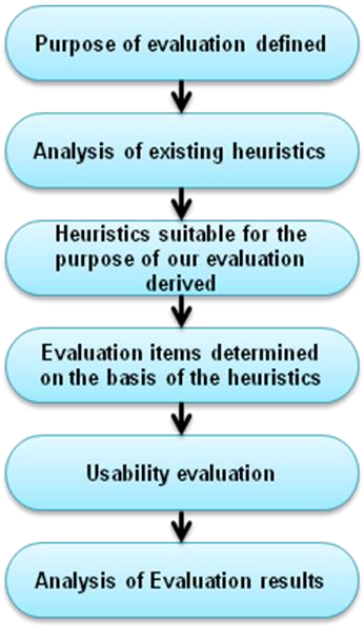


Fig. 11. User evaluation process

Our evaluation heuristics, "ease of use and learnability of operation method", are derived from Jacob Neilsen's 'Ten

Usability Heuristics' which are widely used for evaluating the usability of software and websites and Melissa A .Federoff's 'Game Heuristics'(Federoff, 2002) which studied the usability of game contents.

We quote the following two comments by Jacob Neilsen on "User control and freedom" and "Recognition rather than recall".

TABLE V. CRITERIA FOR USABILITY EVALUATION

1	**User control and freedom**
	Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.
2	**Recognition rather than recall**
	Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

Neilsen's "User control and freedom" is pointing out the need to support "Undo" and "Redo" functions in order that gamers can use the contents freely even after they make mistakes in inputting (Nielson, 1993). Neilsen is recommending here the functions that are not directly related to contents. However, they are important in controlling the contents and using them freely(Federoff, 2002) because if users feel inconvenient or limited in maneuvering the contents, they will lose interest, and this is directly linked to lowering of their interest in the game contents. Discussing the relationship between Ten Usability Heuristics and Game Heuristics, Melissa A. Federoff made the following statements about game interface and game mechanics[Table 5].

TABLE VI. USER CONTROL AND FREEDOM

1	**Game Interface**
	Controls should be customizable and default to industry standard settings
2	**Game Mechanics**
	Feedback should be given immediately to display user control

Melissa A. Federoff's "Recognition rather than recall" emphasizes the importance of intuitive interface which minimizes the amount of information that users have to remember in order to use the interface(Federoff, 2002). The game heuristics related to this are as follows:

TABLE VII. RECOGNITION RATHER THAN RECALL

1	**Game Interface**
	Controls should be intuitive and mapped in a natural way
2	**Game Interface**
	Minimize control options
3	**Game Interface**
	Follow the trends set by the gaming community to shorten the learning curve
4	**Game Interface**
	Do not expect the user to read a manual
5	**Game Mechanics and Play**
	Get the player involved quickly and easily

When the items related to "Ease of use" and "learnability of operation" are put together from "Ten Usability Heuristics" and "Game Heuristics," we learn the following important pieces of information about usability. See Table 7.

TABLE VIII. ITEMS RELATED TO USABILITY EVALUATION IN EXISTING HEURISTICS

Ease of use	a.	User control and freedom
	b.	Feedback should be given immediately to display user control
Learnability	c.	Recognition rather than recall
	d.	Controls should be intuitive and mapped in a natural way
	e.	Minimize control options
	f.	Do not expect the user to read a manual
	g.	Get the player involved quickly and easily

The criteria used in this study for usability evaluation are based on the items listed in Table 7. In the related item part of the table 8, an association item with the table 7 is written.

TABLE IX. CONTENT OF THE QUESTIONNAIRE

Criteria	No.	Content	Related items(Table VII)
Ease of use	1	Is free inputting possible for the control interface that implements each command?	a.
	2	Can users get clear feedback of actions that are on progress according to command?	b.
Learnability	3	Is it easy to learn how to use the control interface of the gamer?	f.
	4	Is it easy to remember the game's control interface?	c.
	5	Can you expect the users who have not learned how to use the commands to use the commands and play the game?	c, d
	6	Is a natural and intuitive operation interface appropriate for inputting commands being applied?	c, d, g.
	7	Is the number of moves required for implementing the commands appropriate?	e.

D. Result of Evaluation

As shown in Fig. 12, the result of usability evaluation indicated that the method proposed in this study gave a higher degree of satisfaction about the feedback on the game in progress because making hand motions in the air (space) is all that users need to do to get the information they want about the positions and interactions. Thus, the feedback on the result of

hand motion in space is delivered more intuitively than the existing multi-touch screen method. However, touching the surface of the touch-screen directly for inputting gave a higher degree of satisfaction to users than touching the air space. This is attributed to the fact that, in general, users have been frequently exposed to the existing method of inputting. This problem can be overcome readily through the learning process of the interface.

TABLE X. RESULT OF USABILITY EVALUATION - EASE OF USE

Characteristics	Item	Content	Proposed system	Existing system
Ease of use	Progress feedback	Is the feedback of the result of inputting clear and correct?	4.48	3.78
	Free input	Is it possible to input easily and freely the operational interface that carries out each command?	4	4.76

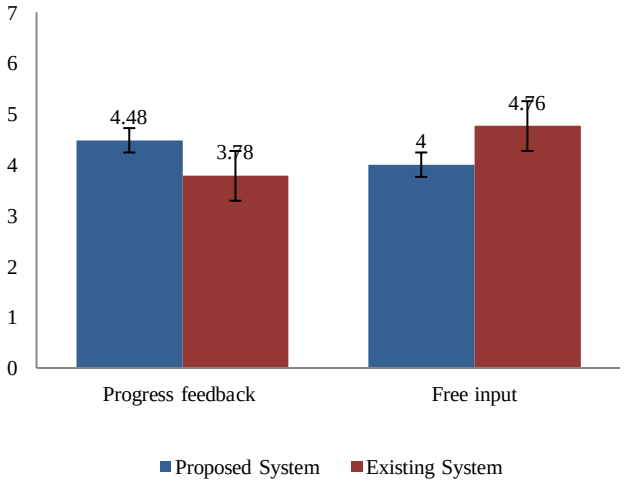


Fig. 12. Result of usability evaluation – ease of use.

As shown in Fig. 13, the result of usability evaluation on the ease of use shows that the proposed method gives a higher degree of satisfaction than the existing method. This means that the space multi-interaction interface is easier and more natural to use than the existing system. For users who are exposed to it for the first time, however, the learnability of the new space interaction interface is somewhat lower than that of the existing multi-touch interaction system. But once users become familiar with the new system after the initial learning process, the functions are easier and more natural to remember and decrease the number of interaction gestures to implement the commands.

TABLE XI. RESULT OF USABILITY EVALUATION - LEARNABILITY OF USE

Subject	Category	Content	Proposed System	Existing System
Learnability of Use	Learnability	Is it easy to learn how to use the interface?	5.96	5.38
	Memory	Is it easy to remember the game interface?	5.7	6.12
	Operative predication	Is it possible to anticipate and be able to use the functions that the users have not learned?	4.34	4.9
	Naturalness	Is a natural and intuitive control interface applied to motions and gestures required to input the command?	4.64	5.6
	Command count	Is the number of starting gestures to implement the command appropriate?	4.94	5.32

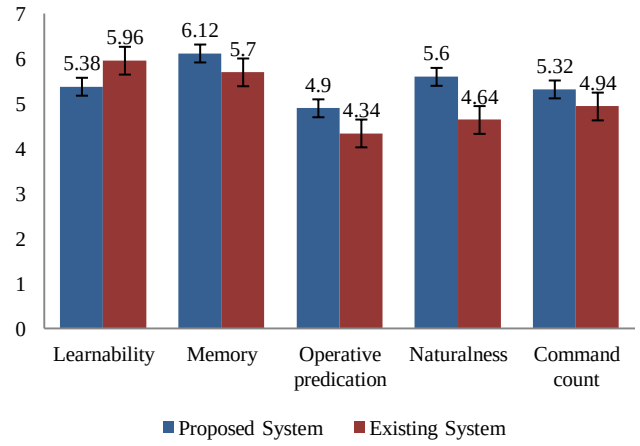


Fig. 13. Result of usability evaluation – Learnability of operation mode

Table 12 lists the result of statistical verification using t-test. The t-test is the statistical method which is necessary when it grasps whether the average difference between the two groups determine whether statistically significant. At the confidence level of 95%, value p is lesser than 0.05 in all items evaluated. Therefore, there is significant difference between the items evaluated. In other words, since the null hypothesis is rejected, there is statistically significant difference between the mean average of the traditional method and that of the proposed method. Statistics and the level (significance level) to compare and judge to reject the null hypothesis when the hypothesis is "statistically significant" are called. In other words, the probability that the result is not enough to think that mere coincidence is meaningful.

TABLE XII. RESULT OF STATISTICAL VERIFICATION

Measures	Proposed system		Existing system		t	p
	Mean	S.D	Mean	S.D		
Progress feedback	4.48	1.403	3.78	1.282	-3.183	0.003
Free input	4	1.385	4.76	1.379	3.040	0.004
Learnability	5.38	1.524	5.96	1.049	3.529	0.001
Memory	6.12	0.872	5.7	1.344	-2.680	0.010
Operative prediction	4.9	1.329	4.34	1.710	-2.527	0.015
Naturalness	5.6	1.355	4.64	1.935	-4.106	0.000
Commands count	5.32	1.377	4.94	1.496	-2.252	0.029

However, in the overall evaluation, the proposed method showed a higher evaluation result than the existing method as shown in Fig. 14. As shown in the experimental result, it is shown from the averages that each result difference is displayed. In the Progress feedback, Memory, Operative prediction, Naturalness, and Commands count items, the proposed system received the higher evaluation but the existing system received the higher evaluation in the Learnabilty and Free input items.

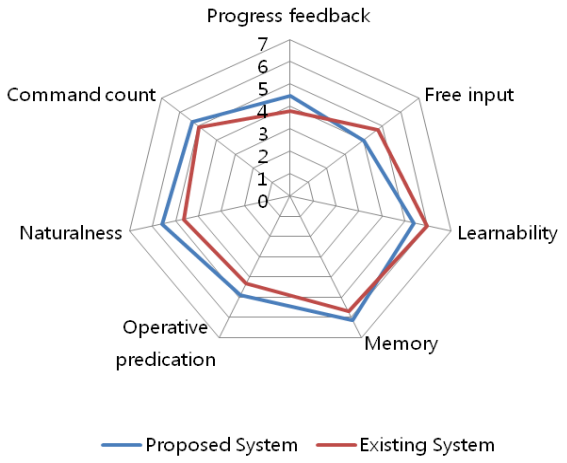


Fig. 14. Scale analysis graph

VI. CONCLUSION

This paper proposes a spatial multi-interaction system to offer an interactive area for multiple users to interact in an ambient display environment. The proposed system creates an interaction surface on which users can interact through the IR-LEDs Array Bar. Thus a variety of interactions are offered. The users are in the interactive space when they stand in arm's length and touch the interaction surface that uses light rays from the IR-LEDs Array Bar. In a ubiquitous ambient environment, the proposed system offers an interactive display system and a user interface method with which users can interact by simple touch motions with the aid of sensing devices for recognition. A performance test indicated that the proposed interface can be as effective as the existing multi-touch screen. With ample explanation on how to operate the system, users can learn various interaction information as they go through the entire operation process, and when the screen size grows larger eventually, multiple users can use the system simultaneously. The system has excellent expandability since it can be operated in a small area as well as a large area without any restrictions, and users can experience many different contents directly by space-touching the interactive surface.

ACKNOWLEDGMENT

This research was financially supported by the Program in the Regional Innovation Center conducted by the Ministry of Trade, Industry and Energy of the Korean Government.

REFERENCES

[1] Aarts, E. (2003). Technological Issues in Ambient Intelligence; Emile Aarts and Stefano Marzano (eds.). *The New Everyday Visions of Ambient Intelligence*, 12-17.

[2] Barrée, R., Chojecki, P., Leiner, U., Müuhlbach, L., & Ruschin, D. (2009). Touchless Interaction-Novel Chances and Challenges. *In Proceedings of the 13th international Conference on Human-Computer Interaction*, Part II: Novel interaction Methods and Techniques (San Diego, CA, July 19 - 24, 2009). Lecture Notes In Computer Science, Vol. 5611, 161-169.

[3] Bellucci, A., Malizia A., Diaz P., & Aedo, I.(2010). Don't touch me: multi-user annotations on a map in large display environments. *AVI '10: Proceedings of the International Conference on Advanced Visual Interfaces*, 391-392.

[4] Baek, S.H., Kim, J.S. & Kim, T.Y. (2005). Infrared LED Tracking and Aim Position Calibration for an Arcade Gun. *Journal of KGS(Korea Game Society)*, 5(1), 3-10.

[5] Choi, W., Kim , T. Y. & Lim, C. S. (2007). A Rehabilitation Training System Using the infrared LED based Motion Analysis. *Journal of KCGS(Korea Computer Graphics Society)*, 13(4), 29-36.

[6] Choi, S. E., Jung, J. W. & Seo, Y. W. (2008). Technology Trend and Application for Tabletop Device and Interactive Wall Display. *Journal of KIISE(Korean Institute of Information Scientists and Engineers)*, 26(3), 5-14.

[7] Federoff, M. A. (2002). Heuristics and Usability Guidelines for the Creation and Evaluation of Fun in Video Games. *Unpublished master's thesis*, Department of Telecommunications of Indiana University.

[8] Hong, D., & Woo, W.(2006). A 3D Vision-Based Ambient User Interface. *International Journal of Human-Computer Interaction*, 1532-7590, Volume 20, Issue 3, 271 – 284.

[9] Hong, S. S., Seo, J. K., Ko, C. S. & Ahn, H. I. (2009). Implementation of Intuitive Method for Controlling Multi-device with Universal Remote Controller. *Proceedings of KHCI 2009*, 646-649.

[10] Holmlid, S. & Björklind, A. (2003). Ambient Intelligence to go, AmiGo white paper on mobile intelligent ambience. Research Report SAR-03-03.

[11] IST Advisory Group. (2003). Ambient intelligence: From vision to reality. Available from ftp://ftp.cordis.lu/pub/ist/docs/istag-ist2003_consolidated_report.pdf.

[12] Johanson, B. (2002). The Interactive Workspaces Project: Experiences with Ubiquitous Computing Rooms. Pervasive Computing Magazine Special Issue on Systems, Vol.2, 67-74.

[13] Kim, Y.M. & Kim J. (2009). IRTS(Infrared Tracking System) for Interactive Multimedia Contents. *Journal of EASKO*, 1(1), 51-59.

[14] Lee, J.C.(2008). Hacking the Nintendo Wii Remote, *IEEE Pervasive Computing*, 7(3), 39-45.

[15] Murata, A.(2006). Eye-gaze input versus mouse: Cursor control as a function of age. *International Journal of Human-Computer Interaction*, 1532-7590, Volume 21, Issue 1, 1 – 14.

[16] Nielsen, J. (1994). Heuristic evaluation. In Nielsen, J., and Mack, R.L. (Eds.), Usability Inspection Methods. *John Wiley & Sons*, New York, NY.

[17] Nielsen, J. (1993). Usability Engineering. *published by Morgan Kaufmann*, San Francisco.

[18] Park, J.S. & Park, J. (2008). LED-bar: An Interaction Tool for Large Display Games. *Journal of KSCG(Korean Society for Computer Game)*, No 15, 49-54.

[19] Russell, D. M., Streitz, N. & Winograd, T. (2005). Building Disappearing Computers. *Communications of the ACM*, 48(3), 42-48.

[20] Stephanidis , C.(2009). Designing for All in Ambient Intelligence Environments: The Interplay of User, Context, and Technology. *International Journal of Human-Computer Interaction*, 1532-7590, Volume 25, Issue 5, 441 – 454.

[21] Streitz, N. A. (2007). From Human-Computer Interaction to Human-Environment Interaction. *ERCIM UI4ALL Ws 2006*, LNCS 4397, 3-13.

[22] Streitz, N., Geißler, J., & Holmer, T. (1998). Roomware for Cooperative Buildings: Integrated Design of Architectural Spaces and Information Spaces. Proceedings of CoBuild '98, Darmstadt, Germany, LNCS Vol.1370, Heidelberg, Germany, Springer, 4-21.

[23] Streitz, N., Tandler, P., Müller-Tomfelde, C., & Konomi, S. (2001). Roomware: Towards the Next Generation of Human-Computer Interaction based on an Integrated Design of Real and Virtual Worlds. *Human-Computer Interaction in the New Millennium*, Addison-Wesley, 553-578.

[24] Tandler, P. (2004). The BEACH application model and software framework for synchronous collaboration in ubiquitous computing environments. *Journal of Systems and Software*, 69(3), 267-296.

[25] Vaughan-Nichols, S. J.(2009). Game-Console Makers Battle over Motion-Sensitive Controllers. Journal Computer, 42(8), 13-15.

[26] Vogel, D. & Balakrishnan, R. (2004). Interactive Public Ambient Displays : Transitioning from Implicit to Explicit Public to Personal, Interaction with Multiple Users. *Proceedings of UIST 2004*, 137-146.

[27] Wisneski, C., Ishii, H., Dahley, A., Gorbet, M., Brave, S., Ullmer, B., & Yarin, P. (1998). Ambient Displays: Turning Architectural Space into an Interface between People and Digital Information. Proceedings of CoBuild '98, Darmstadt, Germany, LNCS Vol.1370, Heidelberg, Germany, Springer, 22-32.

[28] Yoon, J.W., Hong, J.H. & Cho, S.B. (2009). MyWorkspace: VR Platform with an Immersive User Interface. *Proceedings of KHCI 2009*, 52-55.

A particle swarm optimization algorithm for the continuous absolute *p*-center location problem with Euclidean distance

Hassan M. Rabie
PhD Researcher, Decision Support, Faculty of Computers and Information, Cairo University

Dr. Ihab A. El-Khodary
Decision Support, Faculty of Computers and Information, Cairo University

Prof. Assem A. Tharwat
Department of Engineering & Business, Canadian International College

***Abstract*—The *p*-center location problem is concerned with determining the location of *p* centers in a plane/space to serve *n* demand points having fixed locations. The continuous absolute *p*-center location problem attempts to locate facilities anywhere in a space/plane with Euclidean distance. The continuous Euclidean *p*-center location problem seeks to locate *p* facilities so that the maximum Euclidean distance to a set of *n* demand points is minimized. A particle swarm optimization (PSO) algorithm previously advised for the solution of the absolute *p*-center problem on a network has been extended to solve the absolute *p*-center problem on space/plan with Euclidean distance. In this paper we develop a PSO algorithm for the continuous absolute *p*-center location problem to minimize the maximum Euclidean distance from each customer to his/her nearest facility, called "PSO-ED". This problem is proven to be NP-hard. We tested the proposed algorithm "PSO-ED" on a set of 2D and 3D problems and compared the results with a branch and bound algorithm. The numerical experiments show that PSO-ED algorithm can solve optimally location problems with Euclidean distance including up to 1,904,711 points.**

Keywords—absolute p-center; location problem; particle swarm optimization

I. INTRODUCTION

The *p*-center location problem (also called minimax facility location problem) is a major class of location problems. Continuous location problem with Euclidean distance is a main variant of the *p*-center location which is concerned with determining the location of *p* centers in a plane/space to serve *n* demand points having fixed locations. The Euclidean *p*-center location problem seeks to locate a facility so that the maximum Euclidean distance to a set of *n* demand points is minimized [1]. The problem is equivalent to finding the center of the smallest circle enclosing all points [1].

Our main objective is to locate new *p* centers/facilities in the space/plane in such a way that the maximum distance between demand points and their nearest facility becomes minimum. It is assumed that all the facilities are identical and provide the same service to the customers, and there is no limit for the number of customers who can get service from the centers [2]. This kind of location problem is suggested by Hakimi [3, 4], and some of its applications are used to locate fire stations, hospital emergency services, data file location, police stations, and so on. Megiddo and Supowit [5] have shown that the continuous Euclidean *p*-center location problem in the plane is NP-hard, and, such problems are difficult to solve.

James Kennedy and Russell Eberhart [6] in 1995, developed a new metaheuristic algorithm, so-called Particle Swarm Optimization (PSO) algorithm, which is inspired from the flocking of birds, and simulated evolution. Although PSO is comparatively a new metaheuristic algorithm, in various applications; it has been proven to be a robust and efficient tool [7, 8]. PSO has been used mostly to solve continuous optimization problems. The purpose of this paper is to describe a simple put efficient PSO algorithm to solve large-scale Euclidean distance absolute location problem and to test the efficiency of our algorithm. Reported results show that PSO can solve optimally large-scale Euclidean distance *p*-center location problems.

The rest of this paper is organized as follows. Section 2 is devoted to the description of continuous Euclidean *p*-center location problems, while the description of PSO algorithm is given in Section 3. The proposed PSO-ED algorithm for *p*-center problem and the implementation of PSO-ED for solving location problem are explained in Section 4. Section 5 contains experimental results. Section 6 concludes the paper and section 7 contains future work.

II. THE CONTINUOUS EUCLIDEAN *P*-CENTER LOCATION PROBLEM

The minimax location problem seeks to locate a facility so that the maximum distance to a set of demand points is minimized. Therefore, the *p*-center problem involves locating *p* identical facilities to minimize the maximum distance between demand nodes and their closest facilities, i.e. to minimize the worst case possible time spent on the way in providing service. Using Euclidean distances in the plane/space, this problem is equivalent to finding the center of the smallest circle enclosing all points, hence the term "center" regarding this problem [1]. According to [9], usually, the utilization of minimax criterion arises when location of emergency facilities is considered. The facilities will be located in such a way that the response time to the farthest customer will be minimal. Most of the applications arise in emergency service locations such as determining optimal

locations of ambulances, fire stations and police stations where the human life is at stake.

In many cases, the distances between demand and service points are Euclidean [10]. The Euclidean distance location problem seeks to locate p new facilities at some points (x_j, y_j), $j = 1, .., p$ in R^2, within an existing n demand points (a_i, b_i), $i = 1,..., n$. According to [11], the location of p-center facilities in two-dimensional R^2 Euclidean space can be formulated as:

$$\min_{x_j,y_j} \max_i \min_j \left[(a_i - x_j)^2 + (b_i - y_j)^2\right]^{1/2}, j = 1, \dots, p. \quad (1)$$

where (a_i, b_i), $i=1,\dots,n$, are the coordinates of the demand points; (x_j, y_j), $j=1,\dots,p$ are the coordinates of the service facilities (which are to be determined); and $\min_j$ selects for each demand point its closest facility and the $\min_{xj,yj}$ $\max_i$ operations are to be performed.

The continuous absolute p-center location problem with Euclidean distance is an NP-hard problems, even the simplest p-center problems or the approximation to the problem was found to be NP-Hard [12]. The optimal solution can be found in time $O(p^n)$, which is impractical, even for small p and small n [12].

Mainly, there exist two types of the continuous Euclidean p-center location problem, distinguished through the possible location of the service points. The first type, includes facilities which can be located anywhere in the space including the demand points; known as the absolute center location problems. Whereas the second type is the vertex p-center location problems in which facilities can be located only on the demand points. Usually the solution of vertex location problem can be used as an upper bound for the solution for the absolute location problem [13]. In this paper we consider only solving the continuous Euclidean absolute p-center location problem in which each center/facility can be located anywhere in the plane/space including the demand points.

For example, as presented in [9], suppose we need to find 3 centers for the 10 demand points represented as [♦] in blue in Figure 1. Therefore, the continuous Euclidean p-center location problem searches for the optimal location of 3 (p) points (centers) within the problem space in such a way that the maximum distance from these 3 centers to n demand points is minimum than any other 3 points in the space. As Figure 1 shows the location of these 3 centers [○] in red are p_1, p_2 and p_3 with maximum distance equal to 24.0208. Accordingly, the solution vector consists of the coordinates of the 3 center points.

According to [12], many approximation algorithms have been suggested for solving p-center problems. However, recently, we developed a new PSO algorithm for solving the absolute p-center problem on networks, that algorithm has the task to randomly generate swarms on the arcs of network, then, the algorithm has the task to search optimal solution from different combination of swarms [13]. This paper extended that algorithm to solve the absolute p-center location problem on space/plan with Euclidean distance.

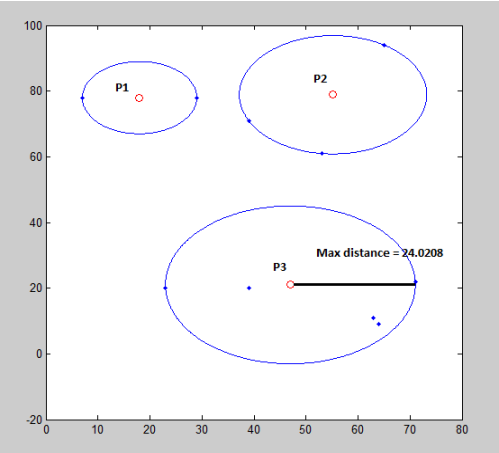


Fig. 1. A set of 10 demand points (in Blue) and service points (in Red) and the maximum distance; from Ref [9].

A. Literature Review

The p-center problem is one of the fundamental problems in the location science. Due to its hardness and importance, it has always been a challenge for the researchers who approached it from different perspectives. When searching for a single center point (p=1); the solution of the problem will be the center of the smallest circle enclosing n given points in the plane [10]. The single center location problem with Euclidean distance was first suggested by James Sylvester in 1814 [1]. Chrystal suggested an algorithm that starts with a large circle that encloses all the points and reduces the radius of the circle iteratively until the smallest circle is obtained [1].

According to [11], the continuous Euclidean p-center location problem was first mentioned in 1958 by Miehle [14] and formulated by Cooper [15] in 1963. Chen suggested a differentiable approximation method to solve the problem [11]. Handler and Mirchandani used relaxation approaches to solve this problem [16]. Daskin [17] presented an optimal algorithm which solves the absolute p-center problem by performing a binary search over possible solution values [17]; the algorithm solved maximal covering sub-problems rather than the set-covering sub-problems solved by Minieka [18].

Recently, Ilhan et al. [19] developed an exact method for solving the vertex location p-center problem in which centers must be chosen only from demand points. The algorithm solved problems with sizes up to 657 points in space. Chen and Chen [10] presented a new relaxation based algorithms for the solution of vertex and absolute continuous p-center problems. The algorithm solved problems with sizes up to 1,817 points in space. Kaveh and Nasr [2] suggested a metaheuristic algorithm called harmony search algorithm. The latter algorithm solved the vertex location problems with sizes up to 4,461 points in space. Calik and Tansel [20] proposed a new integer programming formulation for the p-center problem, in which the optimal p-center solution is obtained by solving a series of simple structured integer programs. The algorithm successfully solved problems with sizes up to 3,038 points in space.

Fayed and Atiya [12] suggested a mixed breadth-depth first strategy to speed up the traversing of the branch and bound tree in order to solve the continuous Euclidean absolute *p*-center location problem. The algorithm was capable of optimally solving problems with size up to 1,904,711 points in space.

According to [10], most of the methods developed for solving the continuous Euclidean problem are geometrical in nature, which involves complex, time consuming search methods for finding the smallest enclosing circle. This includes the repeated solution of relaxed, smaller sub-problems. However, few researches solved the large-scale continuous Euclidean absolute p-center location problems such as [12]. As mentioned before the continuous Euclidean absolute *p*-center problem has been proved to be NP-hard [5] and to approach the *p*-center location problem, we propose a simple algorithm based on PSO (PSO-ED) for solving this problem.

III. Particle Swarm Optimization (PSO)

PSO is a population-based, co-operative search metaheuristic approach introduced in 1995 by Kennedy and Eberhart [6]. PSO inspired from the sociological behavior associated with bird flocking. It is a natural observation that birds can fly in large groups with no collision for extended long distances, making use of their effort to maintain an optimum distance between themselves and their neighbors [21]. PSO was originally used to solve non-linear continuous optimization problems, but more recently it has been used in many practical, real-life application problems [21]. PSO proved to be a successful approach to solve complex continuous problems and is proved to be efficient and robust for solution of combinatorial optimization problems [22].

PSO finds solution for problems that can be represented as a set of points in an *n*-dimensional solution space. PSO is a population-based search algorithm that finds optimal solutions using a set of flying particles with velocities that are dynamically adjusted according to their historical performance, as well as their neighbors in the search space. The population consist form particles which are described as the swarm positions in the *k*-dimensional solution space. Each particle is set into motion through the solution space with a velocity vector representing the particle's speed in each dimension. Each particle has a memory to store its historically best solution (i.e., its best position ever attained in the search space so far, which is also called its experience) [21].

Each particle through flying in the search space generates a solution using directed velocity vector and each particle modifies its velocity to find a better solution (position) by applying its own flying experience (i.e. memory having best position found in the earlier flights) and experience of neighboring particles (i.e. best-found solution of the population). Particles update their positions and velocities as shown below [23]:

- A population of particles is randomly initialized with point positions X_i and velocities Vel_i and a function f is evaluated, using the particle's positional coordinates as input values. Positions and velocities are adjusted and the function evaluated with the new coordinates at each time step.
- When a particle discovers a pattern that is better than any it has found previously, it stores the coordinates in a vector $Pbest_i$.
- The difference between $Pbest_i$ (the best point found by i so far) and the individual's current position is added to the current velocity. Also, the difference between the neighborhood's best position $Gbest_i$ and the individual's current position is also added to its velocity, adjusting it for the next time step. These adjustments to the particle's movement through the space cause it to search around the two best positions.

Following [24], variables X_i and Vel_i are regarded as vectors that show various positions and velocities of particle and in order to find the optimum position of the best position of particle i and its neighbors' best position are recorded as: $Pbest_i$ and $Gbest_i$, respectively. To improve the velocity and position of each particle, the modified velocity and position in the next iteration is calculated as follows:

$$Vel_i^{k+1} = w_k Vel_i^k + c_1 r_1 \left(Pbest_i^k - X_i^k\right) + c_2 r_2 \left(Gbest^k - X_i^k\right) \quad (2)$$

$$X_i^{k+1} = X_i^k + Vel_i^{k+1} \quad (3)$$

where,

- Vel_i^k velocity of particle i at iteration k.
- w_k inertia weight factor which is reduced dynamically to decrease the search area in a gradual fashion. The variable w_k is updated as [22]:

$$w_k = (w_{max} - w_{min}) * \frac{(k_{max} - k)}{k_{max}} + w_{\min},$$

  where, w_{max} and w_{min} denote the maximum and minimum of w_k respectively; k_{max} is a given number of maximum iterations.
- c_1, c_2 acceleration coefficients of the self-recognition component and coefficient of the social component, respectively. The choice of value is $c_1 = c_2 = 2$; and generally referred to as learning factors [7].
- r_1, r_2 random numbers between 0 and 1.
- X_i^k position of particle i at iteration k.
- $Pbest_i^k$ best position of particle i at until iteration k.
- $Gbest_i^k$ best position of the group at until iteration k.

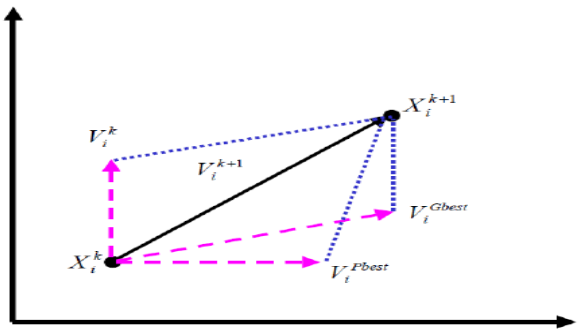


Fig. 2. Updating the position of PSO

IV. PSO-ED FOR *P*-CENTER LOCATION PROBLEM

This paper develops a PSO-ED algorithm for the continuous Euclidean absolute *p*-center location problem which has been proved to be NP-hard. Due to its complexity, hardness and importance it has always been a challenge for researchers who approached it from different perspectives. In [13], we suggested a PSO algorithm for the absolute location problem on networks, in this paper a modified algorithm "PSO-ED" is presented. The main modification is that, we generate the swarm on space/plane limits instead of generating it on arcs of networks as in [13].

PSO-ED algorithm has the task to randomly generate a swarm of birds with size (*Swarm_Size* * *Number_Centers*) for each dimension in *R* within the space/plane which contain the demand points. For each particle in the swarm; we compare the minimax value from each particle – which contains the coordination of centers *p* – to all demand points. Therefore each center will serve a set of demand points i.e. the space/plane will be divided into *p* sections. The procedure is then repeated with the remaining particles in order to find the combination with the best minimum values. The corresponding minimax combination is the optimal location. The PSO-ED search runs in iterations until some predefined stopping criteria is satisfied (*Number of iterations*). The PSO-ED proposed algorithm to solve the continuous Euclidean *p*-center location problem in R^2 can be described as follows:

Step 1. Let *p* represents the number of centers, *s* represents the population size (swarm size, number of particles), and *Vel* represents the swarm velocity.
Step 2. Generate randomly the initial particles positions (x_j, y_j) in R^2 in the range of upper and lower limits for each dimension for each center with size (*s* * *p*). Set the swarm velocities *Vel* to zero.
Step 3. The objective function and fitness value of each particle according to Equation (1) and the *Pbest* is calculated. The best among the *Pbest* is denoted as *Gbest*.
Step 4. The velocity and position of each particle is modified/updated according to Equations (2) and (3), respectively.
Step 5. The objective function of each particle is compared with its *Pbest*. If the current value is better than *Pbest* then *Pbest* value is set equal to the current value and *Pbest* position is set equal to the current position.
Step 6. If the current fitness value is better than the *Gbest*, then update *Gbest* to current best position and fitness value.
Step 7. Steps 4 to 6 are repeated until the maximum number of iterations is met.

V. EXPERIMENTAL RESULTS

For our computational experiments, we applied our algorithm on the 2D and 3D data sets which have been used in [12]. The 2D is from the common TSPLIB library which represents cities/locations in different countries (available at [25, 26]). According to [12], solutions of these problems are very useful in facility location problems where the objective is to minimize the maximum time to reach any location. In TSPLIB instances, the coordinates of points are provided. The number of points range from 7,146 to 1,904,711 points. The 3D geometric model data sets can be found at http://www.ocnus.com/models/, and the number of points ranges from 352 to 437,645.

Fayed and Atiya [12] applied the exact algorithm of branch and bound to large scale location problems on the above mentioned 2D and 3D datasets, with *p* ranging from 3 to 8 centers. They reported comprehensive results of applying the algorithm while trying to achieve accuracies (ϵ) of 10^{-2}, 10^{-3}, and 10^{-4}; the results of which are provided in Table 1. The presented results in Table 1 show, for each problem, the maximum distance between a demand point and its closest center. In our paper, we used the results of $\epsilon = 10^{-2}$ obtained by [12] as an upper bound for our PSO-ED results; presented also in Table 1.

Therefore, Table 1 provides a comparison between using the exact algorithm of branch and bound and the developed PSO-ED algorithm. Comparing the results of the PSO-ED algorithm with those of the branch and bound algorithm at $\epsilon = 10^{-2}$, it is evident that the PSO-ED algorithm provides more accurate results for both the 2D and 3D datasets.

On the other hand, when comparing with the branch and bound with $\epsilon = 10^{-4}$, the bolded values indicate that the PSO-ED slightly outperforms the branch and bound algorithm in many cases for the 2D location problems, and is slightly more accurate for some of the 3D location problems. In either case, even for the instances where the branch and bound outperforms the PSO-ED algorithm, the differences are so small. Accordingly, experimentally we have managed to demonstrate the efficiency of PSO-ED algorithm despite of its simplicity and ease-to-use.

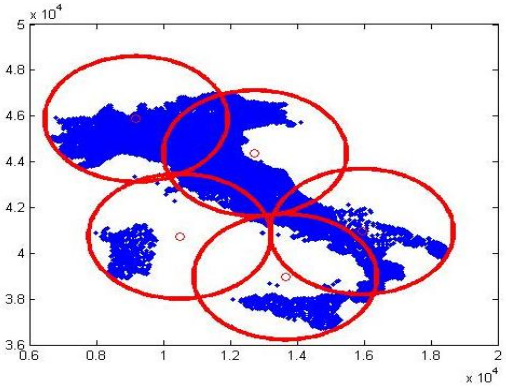
TABLE I. MAXIMUM DISTANCE COMPARISON BETWEEN THE PSO-ED AND BRANCH AND BOUND ALGORITHMS FOR DIFFERENT *P* CENTERS

Problem, Dimension	Demand Points (*n*)	Centers (*p*)	Branch and Bound (max distance)		PSO-ED (max distance)
			10^{-2}	10^{-4}	
Egypt, **2D**	7,146	5	2,363.67	2,352.03	**2,351.90**
		6	2,073.79	2,057.22	**2,056.90**
		7	1,848.18	1,839.08	**1,839.00**
		8	1,722.14	1,713.66	**1,713.60**
USA, **2D**	13,509	5	100,486.74	99,991.80	**99,987.00**
		6	91,950.86	91,310.62	91,411.00
		7	79,917.17	79,566.79	**79,565.00**
		8	75,836.19	75,533.03	75,585.00
Germany, **2D**	15,112	5	2,071.56	2,061.15	**2,061.10**
		6	1,798.07	1,792.68	**1,792.60**
		7	1,647.17	1,645.65	**1,645.60**
		8	1,551.99	1,546.27	1,547.00
Italy, **2D**	16,862	5	2,746.53	2,732.82	2,734.20
		6	2,248.18	2,239.11	**2,239.00**
		7	1,984.57	1,970.92	**1,970.80**
		8	1,880.49	1,860.87	1,878.30

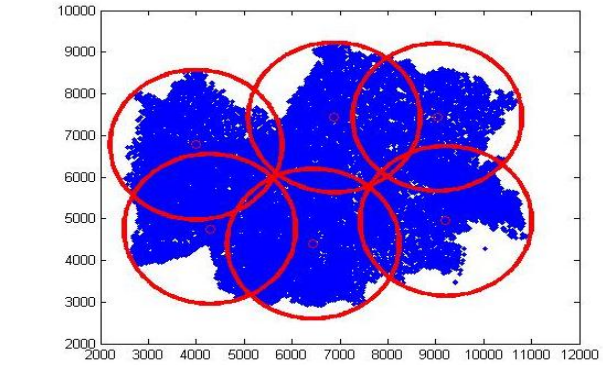
Problem, Dimension	Demand Points (*n*)	Cen-ters (*p*)	Branch and Bound (max distance)		PSO-ED (max distance)
			10^{-2}	10^{-4}	
World, **2D**	1,904,711	5	72.22	71.41	**71.40**
		6	62.55	62.12	62.43
Cat, **3D**	352	3	0.068	0.067	**0.0669**
		4	0.058	0.057	0.0575
		5	0.052	0.051	0.0515
Seashell, **3D**	18,033	3	0.697	0.694	0.6942
		4	0.607	0.605	0.6053
		5	0.549	0.547	0.5485
Bunny, **3D**	35,947	3	0.066	0.065	0.0653
		4	0.056	0.056	**0.0557**
		5	0.051	0.051	**0.0509**
Dragon, **3D**	437,645	3	0.066	0.065	0.0654
		4	0.061	0.060	0.0601

As an example, the location of *p* emergency centers for some countries is illustrated in Figure 3. For the different countries, the cities are demonstrated through the blue points, while the service/emergency centers are the red points. The locations of the *p* centers are optimally selected through the PSO-ED algorithm such that their respective areas encompass all the regions within the optimal maximum distance calculated.

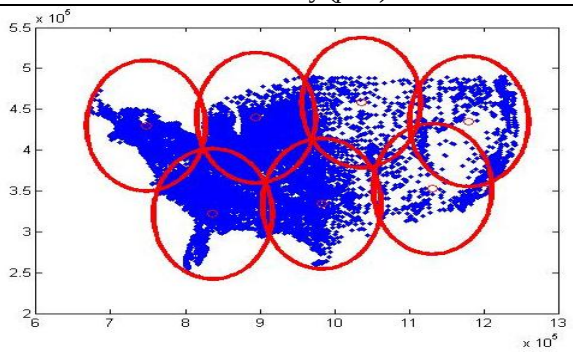
As evident from Figure 3, the service regions of some centers could overlap. For example, the first country presented in Figure 3 is Italy, where the best locations of 5 emergency centers are provided with a maximum coverage distance of 2,734.20.



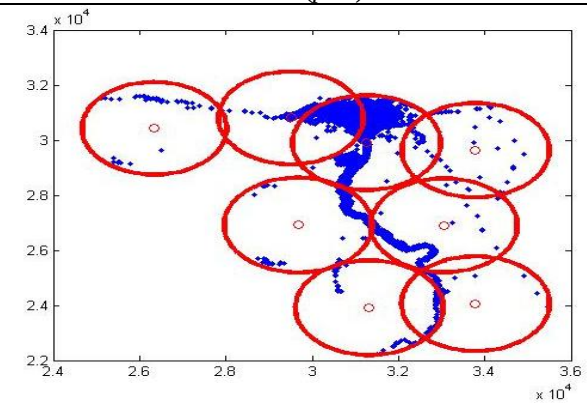
Italy (*p*=5)



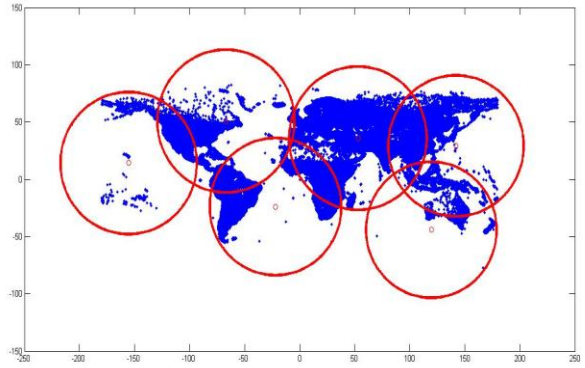
Germany (*p*=6)



USA (*p*=7)



Egypt (*p*=8)



World (*p*=6)

Fig. 3. Location of *p*-center for some countries

VI. CONCLUSION

The continuous absolute *p*-center location problem with Euclidean distance is a complex, and NP-hard problem. Particle swarm optimization (PSO) is a simple and effective algorithm to optimally solve complex continuous problems.

In this paper, a new PSO algorithm for the absolute *p*-center location problem has been developed (PSO-ED). The developed algorithm is simple, easy to apply, and as experimentally shown is an efficient algorithm. PSO-ED has the task to randomly generate a swarm for each dimension in *R* within the space and for each particle; compares the minimax value from each particle. The procedure is then repeated in order to find the combination with the best minimum values.

Results on several well-known test problems are compared with an exact method from the literature. The PSO-ED algorithm used to solve a common 2D and 3D datasets up to 1,904,711 points. We compared our results with a branch and bound algorithm, and the results showed that the PSO-ED algorithm is capable of solving continuous absolute *p*-center location problems optimally.

VII. FUTURE WORK

In this paper, we put forth a new algorithm for the absolute *p*-center location problem that is simple and efficient. The PSO-ED is devised to solve the problem. Although its effectiveness, a hybrid version is recommended in which PSO may be combined with anther metaheuristic technique in order to achieve more accurate results for all instances. We may also expand the usage of the algorithm to solve α-neighbor *p*-center problem in which each demand point is assigned to α service facilities, so that each demand point could withstand the failure of α-1 service facilities.

ACKNOWLEDGMENTS

This work was supported by Al Ezz for Ceramics and Porcelain Company (GEMMA).

REFERENCES

[1] H. A. Eiselt and V. Marianov, Foundations of Location Analysis vol. 155: Springer, 2011.

[2] A. Kaveh and H. Nasr, "Solving the conditional and unconditional p-center problem with modified harmony search: A real case study," Scientia Iranica, vol. 18, pp. 867–877, 2011.

[3] S. L. Hakimi, "Optimum Locations of Switching Centers and the Absolute Centers and Medians of a Graph," Operations Research, vol. 12, pp. 450-459, 1964.

[4] S. L. Hakimi, "Optimum Distribution of Switching Centers in a Communication Network and Some Related Graph Theoretic Problems," Operations Research, vol. 13, pp. 462-475, 1965.

[5] N. megiddo and K. j. Supowits, "On the Complexity Of Some Common Geometric Location Problems," Sofiety for Industrial and Applied Mathematics, vol. 13, pp. 182–196, 1984.

[6] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," Proceedings of IEEE International Conference on Neutral Networks, pp. 1942-1948, 1995.

[7] S. Sumathi and S. Paneerselvam, Computational Intelligence Paradigms: Theory & Applications using MATLAB: Taylor and Francis Group, LLC, 2010.

[8] H. Yapicioglu, A. E. Smith, and G. Dozier, "Solving the semi-desirable facility location problem using bi-objective particle swarm," European Journal of Operational Research, vol. 177, pp. 733–749, 2007.

[9] R. Chen and G. Y. Handler, "Relaxation method for the solution of the minimax location-allocation problem in euclidean space," Naval Research Logistics vol. 34, pp. 775–788, 1987.

[10] D. Chen and R. Chen, "New relaxation-based algorithms for the optimal solution of the continuous and discrete p-center problems," Computers & Operations Research, vol. 36, pp. 1646 -- 1655, 2009.

[11] R. Chen, "Solution of minisum and minimax location–allocation problems with Euclidean distances," Naval Research Logistics Quarterly, vol. 30, pp. 449-459, 1983.

[12] H. A. Fayed and A. F. Atiya, "A mixed breadth-depth first strategy for the branch and bound tree of Euclidean k-center problems," Computational Optimization and Applications, vol. 64, pp. 675-703, 2013.

[13] H. M. Rabie, I. El-Khodary, and A. A. Tharwat, "Applying Particle Swarm Optimization for the Absolute p-center Problem," International Journal of Computer and Information Technology, http://ijcit.com/Vol2Issue5.php, vol. 2, pp. 1010-1015, 2013.

[14] W. Miehle, "Link-Length Minimization in Networks," Operations Research, vol. 6, 1958.

[15] L. Cooper, "Location-Allocation Problems," Operations Research, vol. 11, pp. 331-343, 1963.

[16] G. Handler and P. Mirchandani, Location on networks: Theory and algorithms: MIT Press, Cambridge, 1979.

[17] M. S. Daskin, Network and Discrete Location: Models, Algorithms and Applications: John Wiley and Sons, Inc., New York., 1995.

[18] E. Minieka, "The Centers and Medians of a Graph," Operations Research, vol. 25, pp. 641-650, 1977.

[19] T. Ilhan, F. A. Ozsoy, and M. C. Pinar, "An Effiient Exat Algorithm for the Vertex p-Center Problem and Computational Experiments for Different Set Covering Subproblems," 2002.

[20] H. Calik and B. C. Tansel, "Double bound method for solving the p-center location problem," vol. 40, pp. 2991–2999, 2013.

[21] H. Ahmed and J. Glasgow, "Swarm Intelligence: Concepts, Models and Applications," School of Computing, Queen's University 2012.

[22] P. S. Shelokar, P. Siarry, V. K. Jayaraman, and B. D. Kulkarni, "Particle swarm and ant colony algorithms hybridized for improved continuous optimization," Applied Mathematics and Computation vol. 188, pp. 129–142, 2007.

[23] M. Clerc and J. Kennedy, "The Particle Swarm-Explosion, Stability, and Convergence in a Multidimensional Complex Space," IEEE Transactions on Evolutionary Computation, vol. 6, pp. 58-73, 2002.

[24] M. Eslami, H. Shareef, A. Mohamed, and M. Khajehzadeh, "Optimal Location of PSS Using Improved PSO with Chaotic Sequence," presented at the International Conference on Electrical, Control and Computer Engineering, 2011.

[25] National Traveling Salesman Problems. Available: http://www.math.uwaterloo.ca/tsp/world/countries.html#EG

[26] World TSp. Available: http://www.math.uwaterloo.ca/tsp/world/index.html

Automated Timetabling Using Stochastic Free-Context Grammar Based on Influence-Mapping

Hany Mahgoub
Department of Computer Science
Faculty of Computers and Information
Menoufia University, Shebin El-Kom, EGYPT

Mohamed Altaher
Department of Information Systems
Faculty of Computer and Information Sciences
Ain Shams University, Cairo, EGYPT

***Abstract*—This paper presents a new system that solves the problem of finding suitable class schedule using strongly-typed heuristic search technique. The system is called Automated Timetabling Solver (ATTSolver). The system uses Stochastic Context-Free Grammar rules to build schedule and make use of influence maps to assign the fittest slot (place & time) for each lecture in the timetable. This system is very useful in cases of the need to find valid, diverse, suitable and on-the-fly timetable which takes into account the soft constraints that has been imposed by the user of the system. The performance of the proposed system is compared with the aSc system for the number of tested schedules and the execution time. The results show that the number of tested schedules in the proposed system is always less than that in aSc system. Moreover, the execution time of the proposed system is much better than aSc system in all cases of the sequential runs.**

Keywords*—*Heuristic Search; Automated Timetabling; Stochastic Context-Free Grammar; Influence Map

I. INTRODUCTION

The timetabling problem is a famous problem, consists in scheduling a sequence of lectures among teachers and students in a pre-fixed period of time satisfying a set of constraints of various types [1][2][3][4]. The manual solution of the timetabling problem usually requires many person-days of work. In addition, the solution obtained may be unsatisfactory in some respect; for example a student may not be able to take the courses he/she wants because they are scheduled at the same time. For the above reason, a considerable attention has been devoted to automated timetabling.

The two most common forms of this problem are exam-timetabling problems and course timetabling problems, and in reality, the constraints imposed upon these can often be quite similar. However, the crucial difference between them is usually considered to be that in exam timetables, multiple events can take place in the same room at the same time, whilst in course-timetabling problems; we are generally only allowed one event in a room per timeslot. In automated timetabling, the constraints for both types of timetabling problem generally tend to be separated into two groups: the hard constraints and the soft constraints. Hard constraints have a higher priority than soft, and will usually be mandatory in their satisfaction. Indeed, timetables will usually only be considered feasible if and only if all of the hard constraints of the problem have been satisfied [5]. The Hard constraints and soft constraints are illustrated as follows:

1) Hard Constraints cannot be violated under any circumstances (mainly due to physical restrictions). For example, conflicting lectures (i.e. those which involve common resources such as students) cannot be scheduled simultaneously. Another example is that the number of students taking a lecture cannot exceed the total seating capacity of the rooms.

Soft Constraints are desirable but are not absolutely critical. In practice, it is usually impossible to find feasible solutions that satisfy all of the soft constraints. Soft constraints vary (and sometimes conflict with each other) from one institution to another in terms of both the types and their importance. The most common soft constraint in the timetabling literature is to spread conflicting lectures as much as possible so that students can have enough revision time between exams. An example of another soft constraint which may conflict with this is to schedule all the large exams as early as possible to allow enough time for marking. The quality of timetables is usually measured by checking to what extend the soft constraints are violated in the solutions generated.

One of timetabling problem form is consists of a set of lectures or classes *E* to be scheduled in *N* timeslots (*d* days of *h* hours), a set of rooms *R* in which lectures can take place, a set of students *S* who attend the lectures, and a set of constraints *C* satisfied by rooms and required by lectures [6]. Each student attends a number of lectures and each room has a size. A feasible timetable is one in which all lectures have been assigned a timeslot and a room so that the following hard constraints are satisfied:

1) No student attends more than one lecture at the same time;
2) The room is big enough for all the attending students and satisfies all the constraints required by the lecture;
3) Only one lecture is in each room at any timeslot.
4) In addition, a candidate timetable is penalized equally for each occurrence of the following soft-constraint violations:
5) A student has a class in the last slot of a day;
6) A student has more than two classes in a row;
7) A student has a single class on a day.

Note that the soft constraints have been chosen to be representative of three different classes: the *first* one can be checked with no knowledge of the rest of the timetable; the *second* one can be checked while building a solution, taking

into account the lectures assigned to nearby timeslots; and *finally* the last one can be checked only when the timetable is complete, and all lectures have been assigned a timeslot. The objective of the problem is to minimize the number of soft constraint violations in a feasible solution. All infeasible solutions are considered worthless. In this paper, we introduce a novel technique to solve timetabling problem by making use of Stochastic Context-Free Grammar in conjunction with Influence Map for satisfying soft and hard constraints of the timetabling problem. Furthermore the performance of the approach is compared with the existing one Timetables system like aSc Timetables software.

The rest of this paper is organized as follows: Section II presents the review of literature. Section III presents the proposed approach. Experimental results are presented in section IV. Section V provides conclusion and future work.

II. Review of Literature

In timetabling literature there are many approaches that have been appeared to solve this problem; such as Graph Coloring, Tabu Search, Genetic Algorithm, Artificial Immune Systems and Simulated Annealing Algorithms. Welsh and Powell (1967) represented a very important contribution to the timetabling literature by building the bridge between graph coloring and timetabling, which led to a significant amount of later research on graph heuristics in timetabling [7].

Brailsford, Potts and Smith (1999) introduced various searching methods on constraint satisfaction problems and demonstrated that this technique could be applied to optimization problems [8]. Di Gaspero and Schaerf (2001) carried out a valuable investigation on a family of Tabu Search based techniques whose neighborhoods concerned those which contributed to the violations of hard or soft constraints [9]. Also, White and Xie (2001) developed a four-stage Tabu Search called OTTABU, where solutions were gradually improved by considering more constraints at each stage, for the exam timetabling problem at the University of Ottawa [9]. Abramson (1991) applies simulated annealing to school timetabling [11]. Duong and Lam (2004) employed Simulated Annealing on the initial solutions generated by constraint programming for the exam timetabling problem at HMCM University of Technology [12]. S. Abdullah in (2007) developed a large neighborhood search based on the methodology of improvement graph construction originally developed by Ahuja and Orlin for different optimization problems [13]. Genetic algorithms have been the most studied Evolutionary Algorithms in terms of timetabling research [14].

In particular, hybridizations of genetic algorithms with local search methods (sometimes called memetic algorithms) have led to some success in the field [15]. Also Ulker, Ozcan and Korkmaz (2007) developed a Genetic Algorithm where Linear Linkage Encoding was used as the representation method, different crossover operators were investigated in conjunction with this representation on benchmark graph coloring and exam timetabling problems with hard constraints [16].

Malim, Khader and Mustafa (2006) studied three variants of Artificial Immune systems (a Clonal Selection Algorithm, an Immune Network Algorithms and a Negative Selection Algorithm) and showed that the algorithms can be tailored for both course and exam timetabling problems [17]. Recently, R. Sutar and S. Bichkar (2012) proposed a system uses Genetic Algorithm to design a model for scheduling with challenging constraints considerations [18].

III. The Proposed Approach

In order to make feasible timetable, the aforementioned constraints should be applied by the system which is a heuristic searching technique that use Stochastic Context-Free Grammar (SCFG) parser that produces the entire schedule and selects the correct time and place slot by probabilistic influence mapping. In order to delve in the details we should take the description of the two components of our proposed approach that are the Stochastic Context-Free Grammar and the Influence Mapping.

A. Stochastic Context-Free Grammar (SCFG)

The first component of the system is constructing the schedule using SCFG rules. In order to understand the concept Stochastic Context-Free Grammar, let's describe the concept of Context-Free Grammar. Context-free grammar (CFG) is a collection of context-free phrase structure rules. Each such rule names a constituent type and specifies a possible expansion thereof [19]. These languages are described recursively in terms of each other and primitive Symbols called terminals. The rules relating the variables are called productions. The term context-free comes from the feature that all productions must have a single symbol on its left-hand side, which means that the symbol could always be replaced by the right-hand side of the rule, no matter in what context it occurs.

In general the CFG consists of the following components G = (*S*, *N*, *T*, *R*):

1) A start symbol (S), which is a special non-terminal symbol that appears in the initial sequence generated by the grammar.

2) A set of non-terminal symbols (N), which are placeholders for patterns of terminal symbols that can be generated by the non-terminal symbols.

3) A set of terminal symbols (T), which are the set of rooms, course and time slots generated by the grammar.

4) A set of productions (R), which are rules for replacing (or rewriting) non-terminal symbols (on the left side of the production) in a sequence with other non-terminal or terminal symbols (on the right side of the production).

A Stochastic context-free grammar (SCFG) is a CFG plus a probability distribution on productions: G = (*S*, *N*, *T*, *R*, *P*). The probability of each rule is a conditional probability of choosing a particular right-hand-side to rewrite a given left-hand-side. The Stochastic Context-Free Grammar Parser used here consists of two steps; the *first step* is to prepare the input structure for the actual parser as shown in Fig. 1.

S	<SHEDULE>
N	{ <SHEDULE>, <SLOT>, < LECTURE> , <ROOM, <TIMESLOT >}
T	{ $room_1$-$room_n$, $group_1$- $group_n$, $professor_1$- $professor_n$, $course_1$ – $course_n$,$timeslot_1$- $timeslot_n$}
Rs_{ch1}	< SCHEDULE > → < SLOT>
Rs_{ch2}	< SCHEDULE > → < SLOT> < SCHEDULE >
Rs_{lot}	< SLOT> → {<LECTURE><ROOM><TIMESLOT>}
Rl_{ec}	<LECTURE> → {<COURSE><PROFESSOR><GROUP>}
Rc_1	< COURSE > → $course_1$
Rc_n	< COURSE > → $course_n$
Rp_1	< PROFESSOR > → $professor_1$
Rp_n	< PROFESSOR > → $professor_n$
Rg_1	<GROUP> → $group_1$
Rg_n	<GROUP> → $group_n$
Rr_1	<ROOM> → $room_1$
Rr_n	<ROOM> → $room_n$
Rt_1	<TIMESLOT> → $timeslot_1$
Rt_n	<TIMESLOT> → $timeslot_n$

Fig. 1. The input structure for the Stochastic Context-Free Grammar Parser. The structure is composed of terminals, nonterminals and rules for a timetable.

As Fig. 1 shows, the input structure for the main algorithm composed of a set of terminals and non-terminals.

- The non-terminal <schedule> is taken as the start symbol of parsing.
- The rest of non-terminals are {<SLOT>, < LECTURE>, <ROOM, <TIMESLOT >}. Where the non-terminal < SLOT> represents the lecture in addition to the place and the time, the non-terminal <LECTURE> represents the elements of the lecture which are the course, the groups that take the course, the professor teaching that course. And finally <ROOM> <TIMESLOT > are self-descriptive.

After the previous structure is built, it is used as input for the next stage of the approach. The *second step* is the design of the Context-Free Grammar algorithm as shown in Fig. 2.

- In general, the algorithm applies one of the productions with the start symbol on the left hand size, replacing the start symbol with the right hand side of the production.
- The non-terminal < SHEDULE> is replaced by the rule (Rs_{ch1}) to < SLOT> or to < SLOT>and another < SHEDULE > by the rule (Rs_{ch2}).
- The non-terminal < SLOT> which represents the time slot for a lecture in the schedule is replaced by the rule (Rs_{lot}) into the sequence of non-terminals <LECTURE> <ROOM><TIMESLOT>

```
Algorithm: Context-Free Grammar
----------------------------------------
1:  let S := the start symbol.
2:  let N := a set of non-terminals.
3:  let T := a set of terminals.
5:  let Schedule[ ] := {}empty array of all lectures
6:  let Lecture[ ]={} empty slot
7:  Input: P: productionType
8:  Output: Schedule
9:  let Rules:: = a set of rules that contain P.
10: If TypeOf (P) is Course
11:       let Rules [ ] ← InfluenceMap( )
12: If TypeOf (P) is Professor|| Group|| Room || TimeSlot
13:     let rand:= RandomNumber()%SizeOf(Rules)
14:       let Rules [ ] ← Rules [rand]
15: For each rule in Rules
16:   If rule is nonTerminal
17:       For Each productionType in rule
18:    CFG(productionType)
19:       End for
20:   Else
21:       let Schedule [ ] ←rule.productionType
22: End for
23: Return Schedule
```

Fig. 2. The recursive SCFG parsing algorithm

- The non-terminal <LECTURE> is re-written with the sequence {<COURSE> <PROFESSOR> <GROUP>}; each one of the non-terminals, in the sequence, is re-written to its terminal by selecting its production rule with the higher probability given to that rule to be chosen.
- Lines [1:6] define the storage for the start symbol from which the beginning of the sequence, array of non-terminals , another one for terminals and two empty arrays on for the output schedule and another one for temporary storage for lecture components.
- Line [7] the input for the parser is a productionType which can be any elements of the problem (e.g. professor**,** group, room, timeslot or even a rule that maps non-terminal with another non-terminal or with another terminal).
- Line [8] the output schedule.

- Line [9] define a variable the holds the rules that the input productionType contains.
- The first non-terminal is <COURSE>, which has many rules as the total number of courses in the schedule; the algorithm selects the rule that contains the course returned from the Influence Map algorithm which returns the course with the higher priority this algorithm will be introduced later. the second non-terminal is <PROFESSOR> which has many rules as the total number of professors who gives the course, selected previously, then the professor rule is selected randomly .the last non-terminal is <GROUP> which has the rule that re-write this non-terminal to the group that takes the selected course. this is implemented as following:
- Lines [10:11] the if-condition checks if productionType is the 'Course' element ,if so, then invoke the InfluenceMap procedure to get the course with the highest probability in the form of course-to-availableSlot rule(s).
- Lines [12:14] the if-condition checks if the input productionType is the 'Professor' element (contains the available professors), if so, then selects one of the available professors randomly; this is applied the same for the group, room or timeslot elements.
- Lines [15:22] this is a loop where we iterates over the rules; if the rules of the current productionType are non-terminal (set of professors**,** set of groups, set of rooms or set of timeslot) then call the algorithm recursively against each productionType in that non-terminal; else if the rules are terminal (professor, group, room or timeslot) then add it to the output array.
- Finally, we end up with a sequence of group, professor, course which is the element of the lecture.
- Repeating the process of selecting non-terminal symbols in the sequence, and replacing them with the right hand side of some corresponding production.

B. Influence Maps

In the previous section we constructed the schedule from its elements (professors, groups, course and random slot) using stochastic context free grammar. But, how each production rule determines the probability by which it selects the correct course, professor, etc. in other words how each individual (lecture) is placed in its fittest time and place? For answering this question let's understand and make use of a great concept in the field of artificial intelligence in games, the Influence mapping. Influence mapping is an invaluable and proven game artificial intelligence technique for performing tactical assessment. Influence maps have been used most often in strategy games, but are also useful for many other types of games that require an aspect of tactical analysis [20].

As there is no standard implementation of Influence Map we use simple implementation consists of a grid which has values assigned to each cell based on some function which represents a spatial feature or concept. The Stochastic Influence Mapping Algorithm used in this work is a 2D grid contains the time slots in its column headers and the available days in the headers of the rows representing the time table. In general, every time the algorithm is invoked we fill the grid with zeros the lecture gets its slot (place/time) based on its probability. This means that the lecturer with the higher probability, the higher priority it takes to proceed reserving his slot. The probability value for each lecturer is assigned based on the space available for him, and the less space available to a lecturer increases its probability/priority, and vice versa. Every time the map is constructed we reserve a slot (place/time) for the lecture with the highest probability. The Influence-Mapping algorithm is shown in Fig. 3.

Algorithm: **Influence Map**

1: **let** Courses[] := Array of all courses.
2: **let** ReservedCourses[] := an global array of reserved lectures.
3: **let** tempMaxProb **:= 0**
4: **let** crsWithMaxProb := empty
5: **Input:** Courses
6: **Output:** Lecture with valid room, day and time.
7: **for each** course **in** Courses
8: **if** course **is not in** ReservedCourses
9: **let** nRooms := number of rooms available for the group which the current course belongs to.
10: **let** nDays **:=** number of days available for the professor who gives the course.
11: **let** nTimeSlots **:=** number of time slots per day available for the professor(s) who give(s) the course.
12: **let** nRequiredSlots **:=** number of required time slots for all courses.
13: **let** probabiliy:= nRequiredSlots / (nRooms * nDays * nTimeSlots).
14: **if** probability > tempMaxProb
15: **let** tempMaxProb:= probability
16: **let** crsWithMaxProb:= course
17: **end for**
18: **let** ReservedLecs**[]** ← crsWithMaxProb
19: **Return** crsWithMaxProb.

Fig. 3. The Influence Mapping Algorithm

We can illustrate the work of the Influence Map algorithm as follows:

- Lines [1:4] the array "Courses" represents the array that holds all courses in the problem, the global array

"ReservedCourses" which keeps track of the lectures that has been already selected before, the tempMaxProb and crsWithMaxProb are variables to hold the temporary maximum probability and the course with this maximum.

- Line [5] the input array of all courses.
- Line [6] the output course with highest probability.
- Lines [7:17] the loop iterates over all the courses and checks if the course is not selected before then make simple computation to calculate priority of this course this priority equals the division of the required resources (timeslot and room) by the available resources (timeslot of the available professor and available room)
- Lines [18:19] add the result course to list of the reserved courses and returns.
- Note: The returned type is a productionType this is an abstract type of all the elements in our implementation of the context-free grammar parser.

Usually if we design an algorithm, the best way to present its efficiency is to compute its time complexity. Time complexity is commonly estimated by counting the number of elementary operations performed by the algorithm, where an elementary operation takes a fixed amount of time to perform. Where each time the algorithm is invoked on the same set of courses will iterate over the courses, subtracting the courses that have been already parsed before, that is given by the formula $n*(n+1)/2$; Thus, the time complexity of the Stochastic Influence Mapping algorithm presented here is $O(n^2)$.

C. The Proposed Approach and its Advantages

The solution proposed here to the timetabling problem is a heuristic search method that might not always find the best solution but it is guaranteed to find a good solution in reasonable time. The Stochastic Context Free Grammar Parser takes the structure of the timetable in a suitable format and generates a formal representation of one lecture; each lecture contains professor, course, group, time and room with the constraints imposed by the user. The parser runs as many times as the number of courses in the schedule producing a set of lectures, and each production rules are rewritten based on a specific probability produced by probabilistic Influence Map .

The algorithm should use influence map for lecturers to prioritize the lectures to take the precedence in the assigning slots operation. Sometimes the conflicts occur when all the appropriate slots for a low-probability lecture are reserved for high-probability lecture, to handle such cases; another influence map for slots is constructed. After the parser invokes the influence map for the lecturer to get the lecturer with the highest probability in order to bind his lectures to appropriate and available slots and here comes the role of the influence map for the slots but on the contrary of the lecturer influence map, the slot influence map returned the slot with low-probability. In order to make things simpler we use influence map only for lecturers and whenever the conflicts occur the algorithm re-assign different slot.

One of the advantages of this approach lies in the using of the Context-Free Grammar to construct the time table; using this implies strongly-typed contents to the behavior of the algorithm. Strongly-Typed behavior means specifying one or more restrictions on how operations involving values of different data types can be intermixed. Also, the algorithm gives various, on the fly and feasible schedule each time the algorithm runs. As each time the influence map picks a free slot from the available slots for underlying lecture it does so randomly; thus, every time it assigns different available slot to the lecture.

D. Case Study

This case study illustrates the work of the new approach. Suppose there is a schedule that is composed of four courses as shown in Fig. 4.

SCHEDULE	→	SLOT
SLOT	→	{ LECTURE, ROOM, TIMESLOT }
LECTURE	→	{ GROUP,PROFESSOR,COURSE }
GROUP	→	Diploma
PROFESSOR	→	David | John
COURSE	→	Math | Physics | Calculus | C.Science
ROOM	→	R19
TIMESLOT	→	Sun (8:11)……..Thu(17:20)

Fig. 4. Example of a schedule composed of four courses.

We can apply the re-writing rules of the SCFG timetabling as shown in Fig. 5, where the output is four lectures.

S1	{{ Diploma, Alan, Calculus}, {R19}, Wed (15:18)}
S2	{{ Diploma, John, Physics }, { R19}, Wed (15:18)}
S3	{{ Diploma, David, Eng. Fund. }, { R19}, Sat (8:11)}
S4	{{ Diploma, Alan, Math }, {R19}, Mon (11:14)}

Fig. 5. The output of the re-writing rules is four lectures.

To determine which lecture is chosen the Influence mapping algorithm is applied. The Influence mapping algorithm is giving the precedence to the lecture S2 to reserve the slot R19-Wed (15:18), based on its superior probabilistic value which is indicated by red circle as shown in Fig. 6.

S2	8:11	11:14	15:18
Sun.	0.05555	0	0
Mon.	0	0.05555	0
Wed.	0	0	0.166

Fig. 6. The Influence mapping algorithm is giving the precedence to the lecture S2 based on its superior probabilistic value.

IV. Experimental Results

The performance results for the proposed system are presented by designing and implementing a desktop application program using Java programming language. We called the program "Automated Timetabling Solver (ATTSolver)" and the output is produced in PDF file format.

The implementation of the system is tested with the following specifications:

- Operating system windows 7 ultimate with system type 32-bit
- Hardware specification processor 2 GHz and memory 1.5 GB.
- To evaluate the results and the performance of the ATTSolver system, we compare it with the aSc Timetables software system. aSc timetable generator uses novel in-house developed algorithm. It is loosely based on backtracking with plenty of heuristics and special data structure optimized for maximum performance [21].
- The experiment that is used in the comparison is a timetable consists of one class group with 67 courses and 63 professors and one or more time-constraints on each professor. Fig. 7, 8, and 9 show some of the snapshots screen of the graphical user interface of the ATTSolver system.

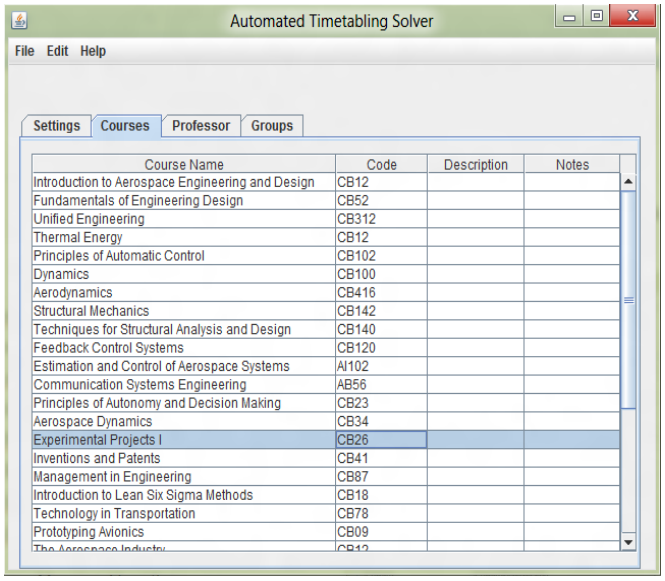


Fig. 7. The form used in the data entry for the information about the courses.

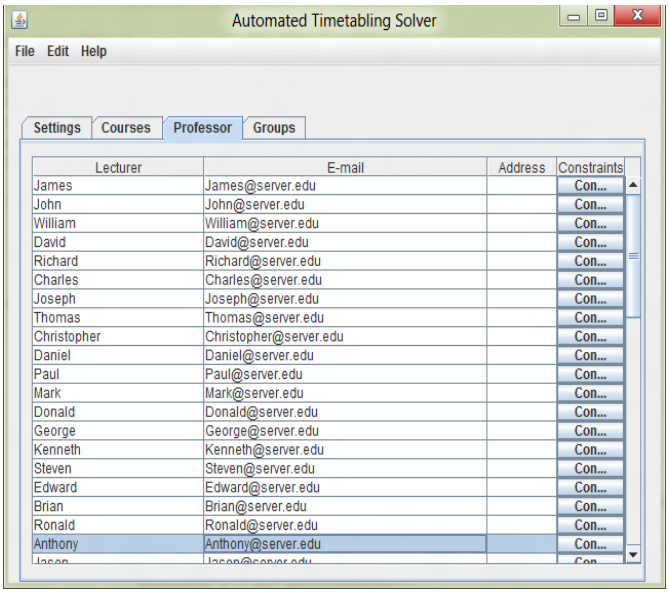


Fig. 8. The form used in the data entry for the information about the professors and their constraints.

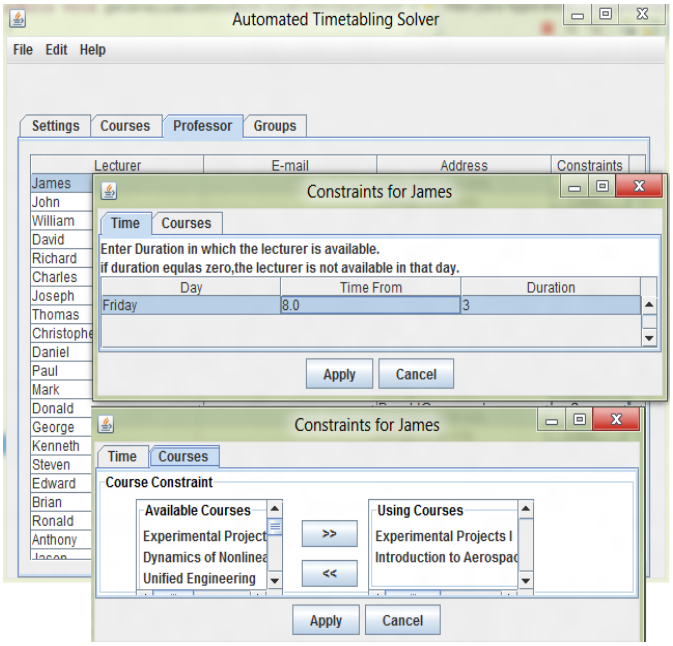


Fig. 9. The dialogs which are responsible for setting constraint for both time and courses assigned for specific professor.

The output of the ATTSolver system is exported in PDF file format as shown in Fig. 10.

	8.0 AM	9.0 AM	10.0 AM	11.0 AM	12.0 PM	1.0 PM	2.0 PM	3.0 PM	4.0 PM	5.0 PM	6.0 PM	7.0 PM
Saturday	DOROTHY Type Theory Room12	LISA Compiler Design Room12	NANCY Programming Languages Room12	KAREN Computer Vision Room12	BRENDA Computer Security Room12	SARAH Robotics Room12	JESSICA Microarchitecture Room12	DEBORAH Digital Logic Room12	SANDRA Pattern Recognition Room12	CYNTHIA Operating Systems Room12	MARY Automata Theory Room12	MARIA Algorithms Room12
Sunday	John Dyanmics of Nonlinear Systems Room12	Charles Compressible Flow Room12	William Inventions and Patents Room12	Smith Principle of Optimal Control Room12	George Aircraft Stability and Control Room12	Daniel Aerodynamics Room12	Brown Computational Methods in Aerospace Engineering Room12	Miller Unified Engineering Room12	Miller Plates and Shells Room12	David Fundamentals of Engineering Design Room12		
Monday	Jones Space System Engineering Room12	Joseph Estimation and Control of Aerospace Systems Room12	Edward Aerospace Dynamics Room12	Paul Technology in Transportation Room12	Mark Structural Mechanics Room12	Paul Thermal Eneregy Room12	Ritchard Prototypin Avionics Room12	Steven Principle of Automatic Control Room12	Thomas Stochastic Estimation and Control Room12	Anthony Dynamic Systems and Control Room12		
Tuesday	Kennet Principles of Autonomy and Decesion Making Room12	James Introduction to Aerospace Engineering and Design Room12	Davis Aerodynamics of Viscous Fluids Room12	Christopher The Aerospace Industry Room12	Kennet Communication Systems Engineering Room12	Jeff Feedback Control Systems Room12	Johnson Techniques for Structural Analysis and Design Room12	James Experimental Projects Room12	Brian Dynamics Room12	Donald Management in Engineering Room12	Jason Computational Mechanics of Matreials Room12	Ronald Introduction to Lean Six Sigma Methods Room12
Wednesday	CAROL Data Mining Room12	PATRICIA Computability Theory Room12	HELEN Image Processing Room12	LAURA Natural Language Processing Room12	MELISSA Computer Networks Room12	SHARON Information retrieval Room12	VIRGINIA Bioinformatics Room12	ANGELA Databases Room12	AMY Ubiquitous Computing Room12	MICHELE Knowledge Representation Room12	JENNIFER Analysis of Algorithms Room12	
Thursday	ELIZABETH Quantum Computing Theory Room12	LINDA Computational ComplexityTheory Room12	RUTH Evolutionary Computation Room12	ANNA Systems Architecture Room12	REBECCA Numerical Analysis Room12	KIMBERLY Medical Image Processing Room12	BETTY Machine Learning Room12	MARGARET Computational Geometry Room12	SHIRLEY Mutiprocessing Room12	DONNA Cognitive Science Room12	SUSAN Data Structures Room12	BARBARA Cryptography Room12
Friday												

Fig. 10. The output table of the ATTSolver system in pdf file format

The experiments are performed to compare the performance of both aSc software system and ATTSolver system for the number of tested schedules, until the fittest solution in each run is satisfied, and the execution time through five sequential runs. The ATTSolver gives best results where the number of tested schedules of the ATTSolver is fewer in compared to the number of tested schedules of the aSc software as shown in Fig. 11. In addition, the execution time of the ATTSolver system is much better than that of the aSc in all cases of the sequential runs. The reason of these results returns to the using of Stochastic Context-Free Grammar rules to build schedule and make use of influence maps to assign the fittest slot (place & time) for each lecture in the timetable.

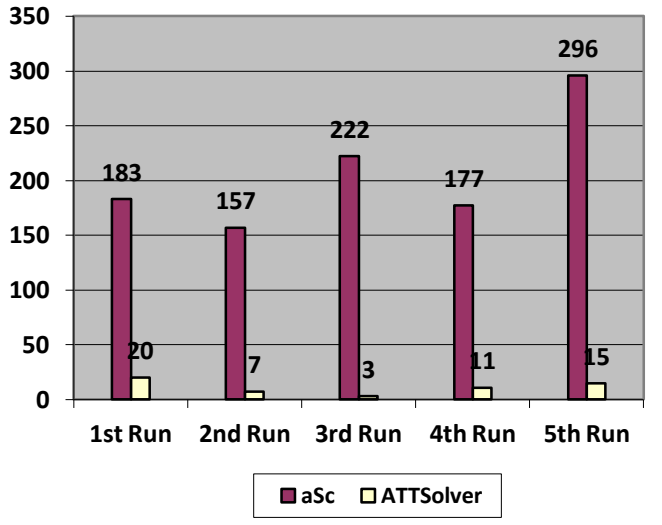


Fig. 11. The number of tested schedules of aSc and ATTSolver

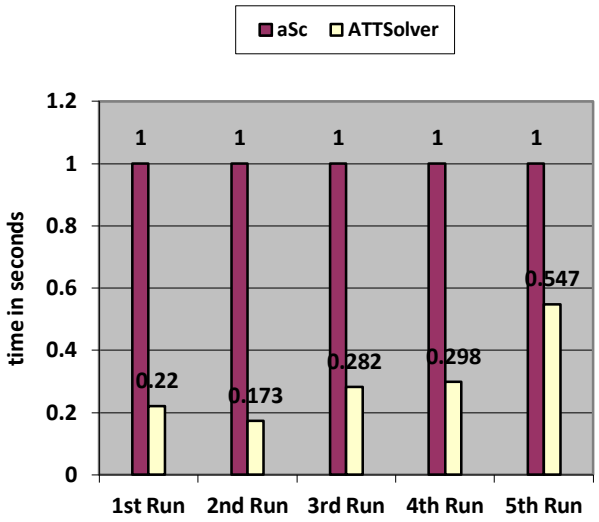


Fig. 12. Execution time of aSc and ATTSolver systems

V. Conclusion and Future Work

This paper presented a solution of the problem of finding suitable class schedule by using strongly-typed heuristic search technique. The system uses Stochastic Context-Free Grammar rules to build schedule and make use of influence maps to assign the fittest slot (place and time) for each lecture in the timetable. The system is taking the soft constraints as well as the hard constraint into the consideration based on its priority making various, instant, and suitable timetable. The results of comparing ATTSolver and aSc systems reveal that the number of tested schedules in ATTSolver system is always less than that in aSc system. Moreover, the execution time for ATTSolver system is much better than that of aSc system in all cases of the sequential runs. In future work we intend to apply the ATTSolver system in the field of software engineering as an approach of refactoring the code and apply the fittest design pattern.

References

[1] A. Schaerf, "A survey of automated timetabling," Artificial Intelligence Review, Kluwer Academic Publishers, vol. 13, pp. 87-127, 1999.

[2] D. Montana, "Strongly typed genetic programming," ACM, vol.3, Issue 2, 1995.

[3] M. Carter and D. Johnson, "Extended clique initialization in examination timetabling," Journal of Operational Research Society, vol. 52, pp. 538-544, 2001.

[4] L. Reis and E. Oliveira, "A language for specifying complete timetabling problems, *Selected Papers from the 3rd International Conference on the Practice and Theory of Automated Timetabling, 2001*.

[5] E. Burke, and S. Petrovic, "Recent research directions in automated timetabling," European Journal of Operational Research, vol. 140, pp. 266–280, 2002..

[6] O. Rossi-Doria, et al. , "A comparison of the performance of different metaheuristics on the timetabling problem," *In Proc. 4th International Conference on the Practice and Theory of Automated Timetabling IV, PATAT02, 2002.*

[7] A. Welsh, and B. Powell, "An upper bound for the chromatic number of a graph and its application to timetabling problems," The Computer Journal, vol. 10, issue 1, pp. 85–86, 1967.

[8] S. Brailsford, C. Potts and B. Smith, "Constraint satisfaction problems: algorithms and applications," ELSEVIER; European Journal of Operational Research,vol. 119, pp. 557-581, 1999.

[9] D. Gaspero, and S. Tabu, "Search techniques for examination timetabling," *Selected Papers from the 3rd International Conference on the Practice and Theory of Automated Timetabling, 2000*.

[10] G. White, B Xie, and S. Zonjic, "Using tabu search with longer-term memory and relaxation to create examination timetables," ELSEVIER, European Journal of Operational Research, vol. 153, Issuo 1 pp. 80-91, Feb. 2004.

[11] D. Abramson, "Constructing school timetables using simulated annealing: sequential and parallel algorithms," Management Science, vol. 37, No. 1, pp. 98-113, Jan. 1991.

[12] T. Duong and K. Lam, "Combining constraint programming and simulated annealing on university exam timetabling," *In Proceedings of the 2nd International Conference in Computer Sciences, Research, Innovation & Vision for the Future (RIVF2004), Hanoi, Vietnam, February 2-5, 2004*.

[13] S. Abdullah, "Heuristic Approaches for university timetabling problems," PhD Thesis, School of Computer Science and Information Technology, University of Nottingham, UK , July 2006.

[14] A. Colorni, M. Dorigo, and V. Maniezzo, "A genetic algorithm to solve the timetable problem," *Technical Report. 90-060 revised, Politecnico di Milano, Italy, 1992.*

[15] Burke and H. Rudova, "Practice and theory of automated timetabling," *Selected Papers from the 6th International Conference. Lecture Notes in Computer Science, vol. 3867, 2007*.

[16] B. Bilgin, E. Ozcan and E. Korkmaz, "An experimental study on hyper-heuristics and exam timetabling," *Proceedings of the 6th International Conference on Practice and Theory of Automated Timetabling, 2006.*

[17] M. Malim, A Khader, A Mustafa, "Artificial immune algorithms for university timetabling," *Proceedings of the 6th International Conference on the Practice & Theory of Automated Timetabling, September, 2006.*

[18] S. Sutar and R. Bichkar, "University timetabling based on hard constraints using genetic algorithm," International Journal of Computer Applications, Vol 42, No.15, March 2012.

[19] C. Antunes and A. Oliveira, "Using context-free grammars to constrain apriori-based algorithms for mining temporal association rules," *Workshop Proceedings on Temporal Data Mining*, 2002.

[20] P. Tozour, "Influence mapping game programming Gemes 2," Chalres River Media, 2001.

[21] http://help.asctimetables.com/index.php

Internet Forensics Framework Based-on Clustering

Imam Riadi
Information Systems Study Program, Faculty of Mathematics and Natural Sciences, Ahmad Dahlan University, Yogyakarta, Indonesia

Jazi Eko Istiyanto, Ahmad Ashari, Subanar
Computer Science Postgraduate Program, Faculty of Mathematics and Natural Sciences, Gadjahmada University, Yogyakarta, Indonesia

***Abstract*—Internet network attacks are complicated and worth studying. The attacks include Denial of Service (DoS). DoS attacks that exploit vulnerabilities found in operating systems, network services and applications. Indicators of DoS attacks, is when legitimate users cannot access the system. This paper proposes a framework for Internet based forensic logs that aims to assist in the investigation process to reveal DoS attacks. The framework in this study consists of several steps, among others : logging into the text file and database as well as identifying an attack based on the packet header length. After the identification process, logs are grouped using k-means clustering algorithm into three levels of attack (dangerous, rather dangerous and not dangerous) based on port numbers and tcpflags of the package. Based on the test results the proposed framework can be grouped into three level attacks and found the attacker with a success rate of 89,02%, so, it can be concluded that the proposed framework can meet the goals set in this research.**

Keywords—framework; forensics; Internet; log; clustering; Denial of Service

I. INTRODUCTION

Background of this research starts from many attacks in the Internet. The attacks such as SYN Flood, IP Spoofing, DoS attacks (Denial of Service), UDP Flood attack, Ping Flood attack, Teardrop attacks, Land Attack, Smurf Attack, Fraggle Attack [1]. DoS attack is a type of computer network attacks that causes the operating system in that server running out of resources. This resulted in the server not being able to serve legitimate user demand and cause the network to be down. Based-on the attacks that often occurs in the Internet network, it is necessary to study forensics to help classifying the log so that attacker information can be immediately known.

Digital forensics is the science dealing with the process of recovery and investigation of material found in digital data, this is often done as part of a criminal investigation [2], [3], [4], in which the scope of digital data comprises a computer system, storage media, electronic documents, or even a sequence of data packets transmitted across computer networks. Network forensics is a part of digital forensics that monitor and analyzes data traffic on the network. Data that are handled in network forensic are dynamic. It is different from that of is digital forensics, where data is static [5].

Research on forensics is related to the data found in network traffic. Network forensics analyzes data traffic through a firewall or intruder detection system in network devices such as routers. The goal is to conduct the traceback to the source of the attack so that the identity of the attacker can be determined [6]. Besides, network forensics has a goal to collect, identify and analyze documents of some processing and transmitting digital data. This activity aims to obtain information or facts related to the attacker [7].

Today's network forensic process particularly by using Internet has increased rapidly. To help facilitate the Internet network forensic process it is needed an alternative solution in the form of a framework to facilitate the discovery of information about the attacker. Framework developed in this study includes overall stages that occur in the forensic process. The whole of forensic process starts from the input in the form of logs obtained from the capture and recording processes in the network traffic. Once the log information is obtained and stored in the database, the log processed using a clustering technique is able to generate the information about attacker needed by the user, in this cases is the investigator. Analysis and log management process requires an additional application that can help implement the framework. Furthermore, that additional application based web is reffered to NFAT (Network Forensic Analysis Tools).

The difference between this study and [8] is [8] focused on framework development, while this paper uses its framework to identify Denial of Service attacks that using NFAT machine. NFAT application is integrated within the framework proposed in this study. To help finding needed information about the attacker, NFAT application needs complex analysis process. To reduce the complexity of data processing, this study utilizes clustering techniques. Clustering technique is one of methods that can be used to facilitate identifying network attacks [9]. Clustering will divide the data into several clusters in which the data in one cluster have similar characteristics and essential equality. The reason of using clustering technique selection in this study is the data characteristic about information of attacker particularly hit access in the numerical form and log information in the network is very large. In addition, to facilitate the process of grouping the log information, it is necessary to facilitate knowing information about attacker in the Internet network. Clustering techniques can be implemented using k-means clustering algorithm.

K-means clustering algorithm is one of the most popular and widely used techniques in the industrial world. K-means clustering method classifies objects in a cluster, the cluster membership value of each cluster centroid is calculated by finding the distance between data and centroid. If the data has the shortest distance from the centroid of a cluster then the data will be the members of the cluster [10]. Web-based applications to detect attacks in the Internet is called NFAT machine (Network Forensic Analysis Tools) that will be used in this study as a proof of concept implementation of a framework for Internet forensic proposed.

II. Current Studies On Network Forensic

Recent research related to this study is divided into two parts, namely the study of forensics in the existing network security and research on clustering techniques often used in data processing.

A. Forensics in Network Security

Several previous studies have been done on digital forensics. In general, the purpose of digital forensic analysis is to identify digital evidence to assist in the investigation. In the mid 1990s the agency guidelines for best practice in the forensic examination of digital technology IOCE (International Organization on Computer Evidence) was established to build the standardization development of digital forensics. The main purpose of IOCE is to combine methods and practices to ensure the ability in using digital evidence. In addition, [11] also developed a scientific working group related to digital evidence for the purpose of forensic guidelines in dealing with digital evidence.

Furthermore [12] presents a technique used in digital forensics to show the methods and tools used for digital forensics. In contrast to digital forensics, few studies done on network forensics has identified several important aspects that are used in network forensics, among others [5] state that network forensic is a part of digital forensic that recently has grown as a very important discipline that used to monitor, especially for the purposes of tracking disorders and attacks. This study suggests it is unlikely that a single tool will be enough for the investigation but it has to use a combination of several tools. While [13] describes a variety of techniques and measures in network security that have been developed to assist the process of digital forensic investigations. It also discusses the network security issues and vulnerabilities that have been exploited by hackers and network intruders. Network preventive measures have been identified through the use of various types of firewall and network architecture. In addition, [14] states the network architecture can have implications on network forensics while the network architecture design is perfect for improving the quality of information produced. According to a statement [15] threats to digital assets has increased so that it is necessary to eliminate the risk from various threats. Attackers have been using anti forensic techniques to hide evidence of Internet crime. Internet forensics equipment should increase the resilience in warding off an ongoing threat. In addition, [16] states that the network forensics solutions based intruder detection analysis need to be followed up in order to record the behavior and analyze the data network intruders in detail . The processing of this data is expected to ensure data integrity and authenticity of data, so the results of data analysis have a good level of credibility in the forensic system.

Forensic process is an activity that combines several disciplines. In contrast to the opinion [17] states that the network forensic analysis is not only a study of science but also need the art to do so. Forensic refers to the use of evidence after the attack to determine how the attack was carried out and what the attacker did. Data traffic on the network is very complicated to be studied. Role of network forensics is to detect abnormal traffic and identify intruders [18]. In addition, there are a few things to watch out where the problem also concerns law enforcement, some of the activities are the process of capturing and analyzing network traffic to get the keywords and information about attacker.

Some tools and techniques analysis used in forensic analysis of network can be seen in table 1 [5].

TABLE I. Some tools used to support network forensics.

Tool	Web Site	Attributes
TCPDump Windump	www.tcpdump.org	F
Ngrep	http://ngrep.sourceforge.net	F
Wireshark	www.wireshark.org	F
Driftnet	www.backtrack-linux.org/backtrack-5-release [Release 3, August 2012]	F
NetworkMiner	www.netresec.com/?page=NetworkMiner	F
Airmon-ng, Airodump-ng, Aireplay-ng, Aircrack-ng	www.backtrack-linux.org/backtrack-5-release [Release 3, August 2012]	F L R C
Kismet	www.kismetwireless.net	F
NetStumbler	www.netstumbler.com	F
Xplico	http://packetstormsecurity.org/files/tags/forensics	F
DeepNines	www.deepnines.com	F
Sleuth Kit	www.sleuthkit.org	F R C
Argus	www.qosient.com/argus	F L
Fenris	http://lcamtuf.coredump.cx/fenris/whatis.shtml	F
Flow-Tools	www.splintered.net/sw/flow-tools	F L
EtherApe	http://etherape.sourceforge.net	F
Honeyd	www.citi.umich.edu/u/provos/honeyd	F
SNORT	www.snort.org	F
Omnipeek /Etherpeek	www.wildpackets.com	F L R
Savant	www.intrusion.com	F R
Forensic Log Analysis GUI	http://sourceforge.net/projects/pyflag	L
Analysis Console for Intrusion Detection	www.andrew.cmu.edu/user/rdanyliw/snort/snorta cid.html	L
Dragon IDS	www.enterasys.com	F R L C
Infinistream	www.netscout.com	F R C
RSA EnVision	www.emc.com/security/rsa-envision.htm	F L R C A
NetDetector	www.niksun.com	F R C A
NetIntercept	www.niksun.com/sandstorm.php	F R C A
NetWitness	www.netwitness.com [www.rsa.com]	F L R C A

Information : F : filter and collect;
L : log analysis;
R : reassembly of data stream;
C : correlation of data;
A : application-layer view.

Some software such as shown in table 1 requires follow-up to the next process could help investigate digital crimes. The investigation of digital crime is indispensable to help the investigation process.

Several attempts to detect such attacks have been carried out using the existing anomaly detection techniques in network traffic using statistical techniques (statistical anomaly detection). This detection technique involves a collection of data related to a legitimate user behavior during a certain time. Currently there are several methods for detecting DoS attacks.

The method is divided into three categories, namely the detection and defense based on the analysis of the protocol characteristics [19], the accumulation [20] and the statistical models on network traffic. Several methods of detection and prevention also have many obstacles, among others : an analysis of the detection and prevention based on the characteristics of the protocol can only be applied to the type of attack that the characteristics of the protocol abnormally occurs [21]. Many types of attacks that do not fit the protocol as well as the accumulation of network traffic statistical models cannot distinguish between normal traffic and large scale [22].

Based on the results of previous studies, carrying out a variety of mechanisms to detect Denial of Service (DoS) has advantages and disadvantages. However, techniques to detect DoS attacks are still complex.

B. Clustering techniques using k-means clustering algorithms

Fundamental issue on software development that supports forensic network is how to determine the appropriate method to facilitate the processing of log data into easily processed data to uncover digital crimes especially those using the Internet as a medium to conduct attacks.

Cluster analysis is the process of analyzing and interpreting a set of data based on similarity. It means that the data is grouped into one cluster due to the same pattern [23]. Clustering includes the type of learning that is unsupervised. Supervised learning and unsupervised learning have a different way of working which is very significant. To have some kinds of unsupervised learning algorithms, [10] has studied the comparison against some types of clustering algorithms. In doing this comparison [10] used several parameters such as the popularity, versatility and easily applied to the data in bulk. There are four kinds of clustering algorithms compared to performance, such as: k-means, hiearchical clustering, self organization map (SOM) and the expectation maximization (EM Clustering). Based on the test results can be concluded that the performance of the k-means algorithm and the EM is better than the hiearchical clustering algorithm. In general, partitioning algorithms such as k-means and EM are highly recommended to be used in the large size of data. It is different from a hierarchical clustering algorithms that have good performance when they are used on small data size.

Furthermore [24] describes an intruder detection system that has been developed to achieve high efficiency and improve the accuracy of detection and classification. The proposed system consists of two stages. The first stage is to detect the attack and the second stage is for classification of attacks. Data mining techniques can be used to improve the detection rate and reduce the false alarm rate. In addition [25] states that the k-means algorithm is needed to determine the final number of clusters (k) before. Based on the results obtained by k-means algorithm turns out better than the FCM algorithm. FCM produces result which is close to the k-means clustering, but it still takes longer than computing k-means. K-means algorithm appears to be superior compared with the Fuzzy C-Means algorithm (FCM).

III. NETWORK FORENSICS

Network forensics is an attempt to find the attacker information to look for potential evidence after an attack or incident. These attacks include probing, DoS, user to root (U2R) and remote to local. Network forensics is the process of capturing, annotating and analyzing network activity in order to find digital evidence of an attack or crime committed using a computer network so that offenders can be prosecuted under applicable laws as illustrated in Fig.1 [26]. Digital evidence can be identified from the recognized attack patterns, deviations from normal behavior or deviations from the network security policy that is applied to the network. Network forensics has a variety of activities and techniques of analysis, such as: analysis of existing processes in the IDS , the analysis of network traffic [11] and the analysis of network device itself [27], all considered the part of a network forensics.

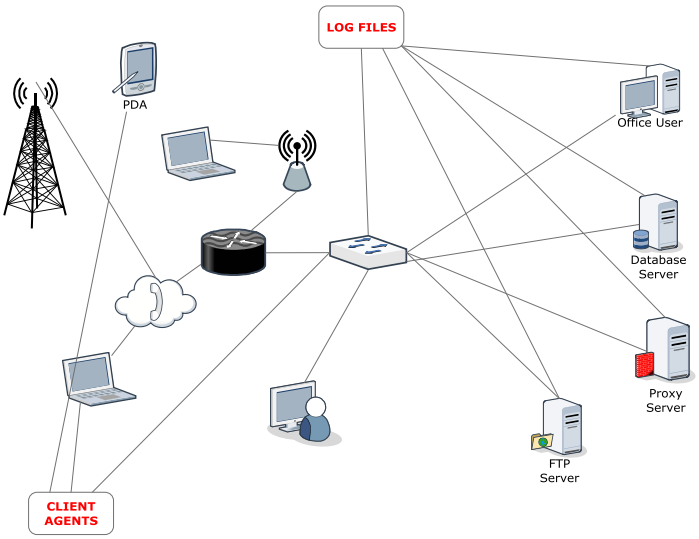


Fig. 1. Overview of network forensics process

Digital evidence can be gathered from various sources depending on the needs and changes in the investigation. Digital evidence can be collected at the server level, the level of proxy or some other sources. For example, at the server level digital evidence can be gathered from web server logs that store browsing activity behavior frequented. The log describes the user that accesses the website and what it does. Several sources includes the contents of the device and the network traffic through both wired and wireless networks. For example, digital evidence can be gathered from the data extracted by the packet sniffer such as tcpdump [28] to monitor incoming traffic in the network. Currently, the number of criminal evidence in the computer continues to increase, even most of the evidence is still used to represent the traditional or conventional crime.

A. Level Attacks in Computer Networks

Computer security often focuses on preventing attacks using authentication, filtering and encryption techniques. Nevertheles another important aspect is the act of detecting attacks after the violation occurs in a network attack.

There are two general approaches to determine whether there is an attack in a network such as digital signature detection, where the pattern of attacks signal will be searched as well as anomaly detection, where abnormal deviations in the network will be detected to determine whether there is an attack or not. Deviations will be divided into several attack types. Table 2 shows the grouping of several types of attacks based on the level of attacks [29].

TABLE II. LEVEL OF ATTACKS IN COMPUTER NETWORKS

No	Level of Attacks	Port / Protocol	TCP Flags	Information
1	Dangerous	80 / TCP	16,32	HTTP
		8080 / TCP	16,32	HTTP alternate
		443 / TCP	16,32	HTTPS (Hypertext Transfer Protocol over SSL/TLS)
		20 / TCP	16,32	FTP data transfer
		21 / TCP	16,32	FTP control (command)
		22 / TCP	16,32	SSH
		23 / TCP	16,32	Telnet protocol
		53 / UDP	-	DNS
2	Rather Dangerous	161 / TCP	20 - 24	SNMP
		143 / TCP	20 - 24	IMAP
		162 / TCP	20 - 24	SNMPTRAP
		110 / TCP	20 - 24	POP3
		993 / TCP	20 - 24	IMAPS
		137 / UDP	-	NetBIOS
		161 / UDP	-	SNMP
3	Not Dangerous	In addition to the above mentioned	TCP (20-27) UDP (-)	In addition to the above mentioned

The attack happens in the computer refers to the protocols and ports used. Based on the protocol and the port level, attacks will be grouped into three levels consisting of malicious port 80, 8080, 443, 20, 21, 22, 23, 53, and somewhat dangerous level consisting of ports 161, 143, 162, 110, 993, 137, 161 and harmless level consisting of a port in addition to those at the previous level.

B. Attack Detection Using IP Header

Threat of attacks in computer networks has grown rapidly. Monitoring and analysis of packet data traffic in the network is done by examining all packet headers and threats to each package. Normal pack behavior was analyzed according to each protocol and header. Each protocol has a header and function in accordance with the protocol TCP/IP [30].

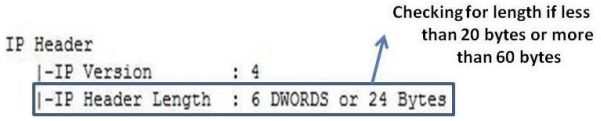


Fig. 2. Field packets that have been observed and analyzed

Fig. 2 shows that IP hearder lenght in the IPV4 must be equal or above 20 bytes and equal to or below 60 bytes. If the IP header length is less than 20 bytes or 60 bytes above, it can be suspected that there is an attack contained in a computer network.

This study has a case study about the types of attacks that often lead to network services disrupted. The disruption is caused by attackers who launched an action by sending and flooding data packets in the Internet network. All data traffic on the network can be saved into the log. This log is very important to help identifying the cause of the attack and can be used as evidence for the investigation. Logs can be obtained in one way to capture data traffic on an interface that is connected to a peripheral or router. Capture traffic activity on the network can be done using several tools, one of which can be used is tcpdump. Tcpdump application has the ability to capture data traffic on a particular interface that is specified. Results of the tcpdump application are then saved as a log that later can be used to reconstruct a digital crime using the Internet.

In this research, a case study used to process forensic is DoS attacks (Denial of Service). The DoS attacks include malicious attacks and often causes the system or network slow even down. DoS attacks examined in this study is a DoS attack that attacks port 80 (http), port 443 (https), port 21 (ftp) and port 22 (ssh). Reason for selecting the port for the case study because the service mentioned above is a public service which is often used by users to utilize the Internet network.

IV. FRAMEWORK FOR INTERNET FORENSICS

This section discusses the identification of critical needs in Internet forensics based on digital crime case studies as described above. NFAT engine development (Network Forensic Analysis Tools) proposed in this study requires a supporting infrastructure consisting of multiple hardware requirements (hardware) and some software (software) supports.

A. Proposed Framework for Internet Forensics

This section discusses a framework proposed in the study. This framework was developed based on the identification of the requirements needed in Internet forensics. Stages identified in Internet forensics are shown in Fig. 3 [8].

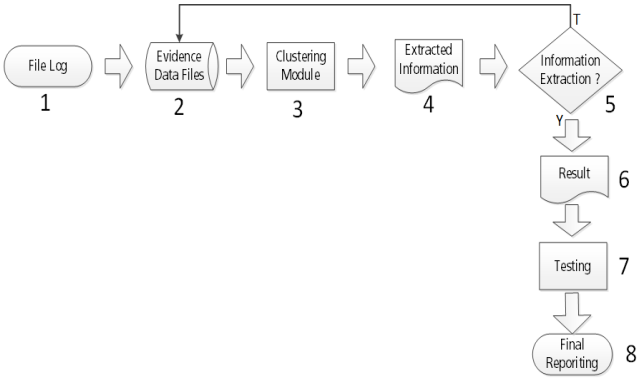


Fig. 3. Proposed framework for Internet forensics

Complete systematic of Internet forensics framework proposed in this research includes several stages. There are 8 stages of the process that must be performed sequentially. Details of each stage are shown in Table 3.

TABLE III. DETAILS OF INTERNET NETWORK FORENSICS PROCESS STAGES

No	Process	Information
1	File *Log*	This log is generated from tcpdump output, at this stage in realtime tcpdump application will capture all the data packets passing through the network interface specified.
2	Evidence data files	Evidence in question here is the original log, the output of the tcpdump application that is stored in a text file.
3	Clustering Module	NFAT module developed in this research is the application modules that can classify the level of these types of attacks into 3 groups (dangerous, rather dangerous and not dangerous). The concept is applied in this module uses clustering techniques using k-means algorithm.
4	Extracted Information	At this stage the log is already saved in the database to extract the data in accordance with the purpose of investigation. Log in information is stored in a database NFAT tools that apply the concept of partitioning MySQL database using horizontal partitioning techniques.
5	Confirmation Extract Information	At this stage, the investigator will conduct relevant confirmation log that contains the IP address is generated by the clustering module.
6	Result	At this stage, detailed information will be obtained IP addresses that have been identified through clustering module. With the help of application services that can be accessed via the URL: http://www.robtex.com the detail information of IP addresses that have been found will be clearer information relevant ASN (autonomous System Number) is used.
7	Testing	At this stage, the test will be held as a proof of concept dati form the framework for the proposed Internet forensics, testing is done using the software LOIC and DoSHTTP which will perform the simulation engine to NFAT and test results will be verified by the original logs in the form of a text file.
8	Final Re*port*ing	At this stage the information is already known to the attacker can be molded according to the needs of the investigator.

B. Architectural Design Network

This section discusses the design of network architecture used to implement the NFAT machine. The approach used in the implementation of network topology uses hierarchy concept. In a hierarchical network concept the network is divided into several parts according to the functions and services provided at each proficiency level layer. Design of network architecture used in this study is shown in Fig. 4.

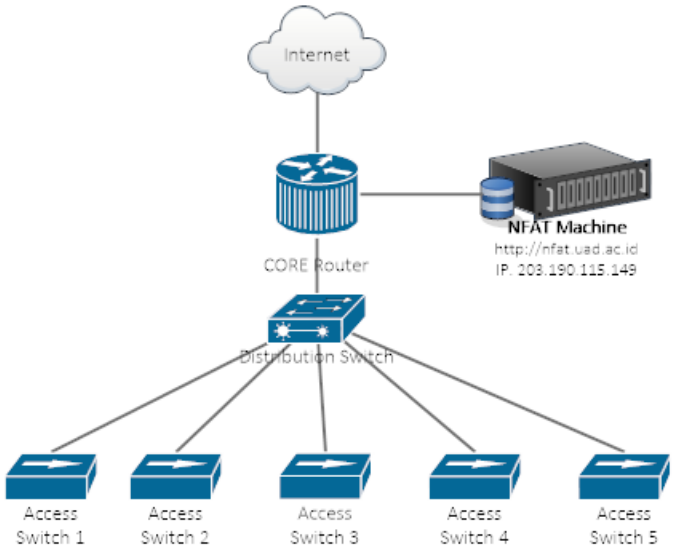


Fig. 4. NFAT machine network architecture design.

C. Implementation of Data Traffic Arrest

The purpose of this Internet forensics is to help finding information about attacker for digital crime commited in the Internet network. It requires careful analysis process so that the goal of the forensic process can be achieved. The process of arrest data traffic in computer network is done using tcpdump assistive software. The function of this application is capturing data traffic in real time and saving it as a log. That tcpdump application arrests all data traffic passing through the interface eth0 and displays it in a time format that is suitable with timestamp in the form of IP Address and display information and protocol header of the data packet. then, The results of the tcpdump application is then stored in the form of a text file with the name of NFAT-eth0.log. Log in the form of a text file then stored and used as the original log verification if it is required by the investigator. In addition, log tcpdump output result is also stored into the database. Logs derived from the process of arrest data traffic is parsed using regex so that logs can be saved to the database by fields (timestamp, source mac address, mac destination address, source address, source port, destination address, destination port, and protocol length).

D. Implementation of Data Grouping

The process after storing log file in the databases is developing NFAT machine and grouping the data to find information about the attacker using Internet network. This study uses clustering techniques to group logs that have been stored in the database.

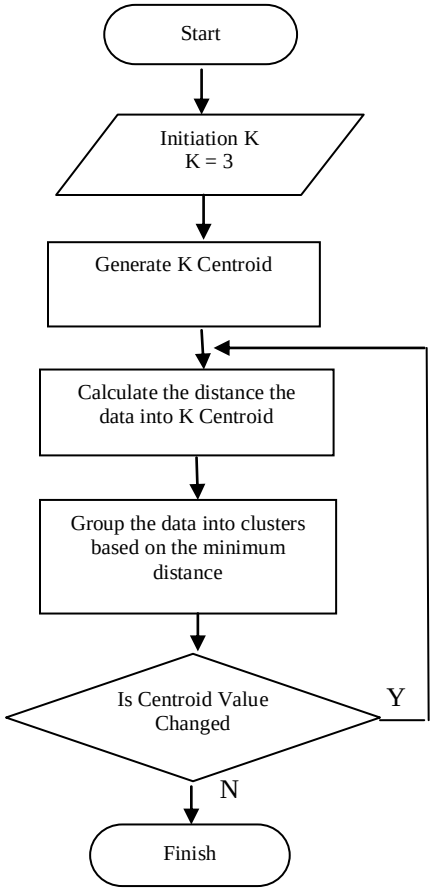


Fig. 5. Flowchart k-means clustering algorithm for grouping log.

NFAT engine that was developed in this study uses the k-means clustering algorithm. Clustering process is used in order to help finding information by classifying the attacker logs into three groups attack levels, namely: dangerous attack, rather dangerous attacks, and not dangerous attacks. Detail steps of process that occur in the process of grouping data using k-means clustering algorithm can be seen in the flowchart in Fig. 5.

E. Database Implementation

Result of the data traffic capture process on the network then is stored in NFAT machine database. Database server used to store logs has been previously filtered using MySQL. NFAT machine has 4 tables that serve to capture and process the results of data traffic logs to facilitate the process of storage and retrieval of information about the attacker as shown in Fig. 6.

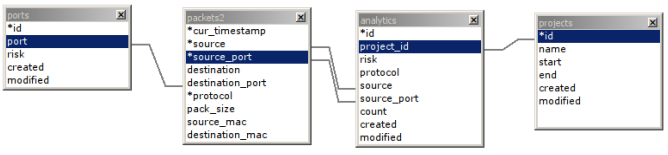


Fig. 6. NFAT engine database schema diagram

V. The Result Analysis NFAT Machine

The results are divided into two stages, The first stage, NFAT engine captures existing packet of data traffic in the network that next will be carried out to devide the clustering process into three levels of attack. The next stage, NFAT machine optimizes the search process and storage of logs into the database. Follows are details of the stages done by NFAT machine.

The first stage of the forensic process starts from collecting information related to the user reports to the investigators then followed by managing the information sought by the data and time attack events. In the analysis phase, the results of the data traffic on the network will be saved in the original logs in the form of a text file and also stored in the database. Incidents of attacks are captured and stored by the NFAT (Network Forensic Analysis Tools) machine. Information needed by investigators will be extracted from the clustering module, where the profile creation process and the analysis time are used as part of the incident investigation process. The resulting interim results clustering module will be verified by the investigator. If there is a verification process need to be clarified about the IP address information that has been generated by the clustering module, investigators can then re-check NFAT into the engine to make sure that the IP address is an IP Address of the suspected assailants who had attacked the system through the Internet network. It can be assisted and linked to the previous stage to repair information, whether the information was sufficient or not. In the final stages of reporting, information related to the attacker who has been found can be used to help uncover digital crimes committed using the Internet network.

Clustering module works using k-means algorithm, where the module can perform an attack level grouping into three groups:

1) dangerous attack,
2) rather dangerous attack,
3) not dangerous attack.

Based on the data stored in the database log, the clustering process will be carried out through the following steps.

1) Specify value of k as the number of clusters to be formed.
2) Generate initial k centroids randomly.
3) Calculate the distance of each data to each centroid.
4) Data will flock around the nearest centroid.
5) Determine the new centroid positions by calculating the average values of the data from the same centroid.
6) Go to step 3 if the new centroid position are not the same.

Results for data clustering, because of its random nature, is highly dependent on centroid generation, this is cause the result of attack detection on the data is always changing. After the attack data clustering process is done, then every cluster results do cluster labeling are included in the dangerous, rather dangerous or not dangerous level of attacks. After cluster labeling, the data entered are checked for the next process of grouping noted in the report. The process of clustering using k-means algorithm is shown in Fig. 7.

N

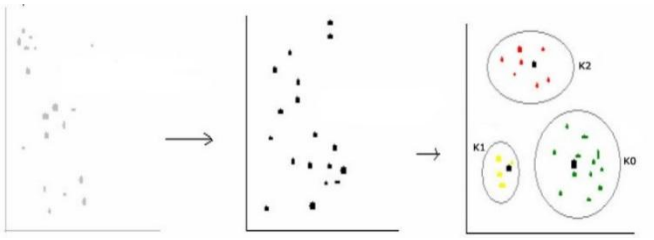


Fig. 7. The process of data clustering with k-means attack

Based on Fig. 7 clusters formed above are the best clusters obtained from the clusters that has the smallest value variants. From the above formed clusters, each cluster for the data is already formed but not yet labeled. At labelling process from the largest to the smallest variants, the result show that clusters K0, K1 clusters and clusters K2, K0 are not dangerous attack, cluster K1 is rather dangerous attack and cluster K2 is dangerous attack.

Logs generated by NFAT machine consist of several items, including IP address, port and hit. Example of a successful log data stored by NFAT machines is shown in Table 4.

TABLE IV. LOG DATA IS STORED BY NFAT MACHINES

No	IP Address	Port	Hit
1	103.19.183.67	80	90
2	202.67.40.24	80	50
3	192.168.100.15	80	30
4	202.67.40.25	80	70
5	203.190.115.149	80	80
6	202.67.40.11	80	40
7	103.19.180.2	80	20
8	203.190.112.231	80	30
9	202.67.40.5	80	50
10	198.24.130.167	80	60

The next stage is the clustering process which is based on the log data in Table 3. where the clustering parameter used is the number of IP Address that are grouped into three categories based on the number of hits. Those are : dangerous attacks, rather dangerous attacks and not dangerous attack. The algorithm used to perform the data clustering is k-means clustering using euclidean distance concept. Based on the results, the calculation is shown in Table 5.

TABLE V. DETAILS OF FINAL RESULTS FOR THE TENTH STAGE OF DATA CLUSTERING.

No	IP Address	Hit	Cluster	Category
1	103.19.183.67	90	1	dangerous attack
2	203.190.115.149	80	1	dangerous attack
3	202.67.40.25	70	1	dangerous attack
4	198.24.130.167	60	1	dangerous attack
5	202.67.40.24	50	2	rather dangerous attack
6	202.67.40.5	50	2	rather dangerous attack
7	202.67.40.11	40	2	rather dangerous attack
8	192.168.100.15	30	3	not dangerous attack
9	203.190.112.231	30	3	not dangerous attack
10	103.19.180.2	20	3	not dangerous attack

Implementation of the clustering process in NFAT engine was developed using the PHP programming language using CakePHP framework presented in the form of display. That kind of level of attack can be viewed in more detail by the port that will be analyzed. In addition, the clustering process is also displayed in graphical information by categorizing type level of attacks where dangerous attack is colored red, rather dangerous attack are colored yellow and not dangerous attack is colored green as presented in Fig. 8.

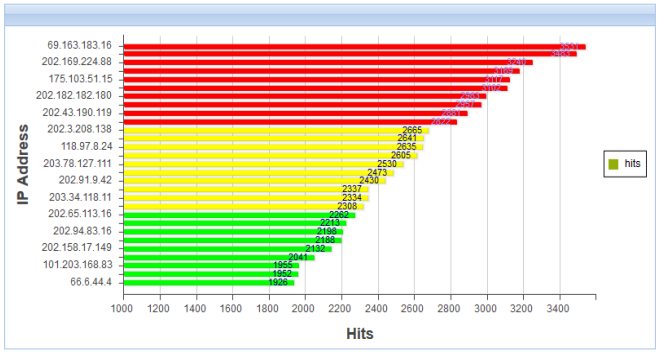


Fig. 8. The result of the clustering process step NFAT machine.

B. Scenario Testing

This section discusses what needs to be prepared prior to testing. Understanding the needs and design testing is important that the testing phase goes well according to plan. This study develops a framework forensics through the development of Internet and tests the system and field experiments in the form of test scenarios. Scenario testing will be done with the topology shown in Fig. 9.

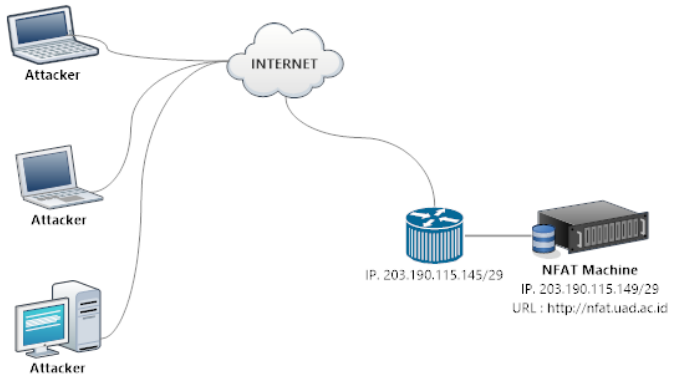


Fig. 9. NFAT machine design test scenarios

Fig. 9 explains that the machine has NFAT domain URL with an IP Address 203 190 115 149 nfat.uad.ac.id connected to the Internet via a router. In addition, there are some attackers who carried out the attack through the Internet network. This attack uses testing tools DoSHTTP. DoSHTTP software downloaded from the site (http://www.socketsoft.net/). It is a tool used to simulate an attack into NFAT machine. Tool DoSHTTP is a testing tool for HTTP flood DoS requests which sends packets to a NFAT machine. Ports that can be simulated using the tools DoSHTTP is port 80 (http) only. The testing process is done by inserting attacks IP address or domain of the NFAT machine nfat.uad.ac.id 203 190 115 149 or domains as shown in Fig. 10.

Fig. 10. Software testing machine DoSHTTP for NFAT

C. DoS Attack Scenario Port 80 (http) Using DoSHTTP

Attacker use this software to simulate an attack on a target machine that has an IP Address 203 190 115 149. IP addresses are included in the target URL or can be replaced with the domain name of the target machine (nfat.uad.ac.id) to be attacked. User Agent on DoSHTTP applications to select the type of browser will be used to simulate the attack. Sockets on DoSHTTP application show the magnitude of the package to be delivered to the target machine. Start Flood used to launch an attack on the target machine.

Results of attack scenarios using software DoSHTTP is then recorded especially in the request packet that will be checked with result of log that is stored in the text file form in the database. Based on the results of testing scenarios conducted a number of attacks ten times using a different IP address, obtained the results as shown in Table 6.

TABLE VI. RESULTS OF TESTING SCENARIOS USING PORT 80 ATTACKS DOSHTTP

No	IP Address Attacker	Port	Request Issued (Hit)	Request Received (Hit)	Detect (%)
1	103.19.183.67	80	100489	89434	88,99
2	202.67.40.24	80	100213	89217	89,03
3	192.168.100.15	80	100119	89118	89,01
4	202.67.40.25	80	100247	89232	89,01
5	203.190.115.149	80	100320	89290	89,01
6	202.67.40.11	80	100393	89360	89,01
7	103.19.180.2	80	100474	89469	89,05
8	203.190.112.231	80	100297	89286	89,02
9	202.67.40.5	80	100317	89304	89,02
10	198.24.130.167	80	100424	89415	89,04

Based on the test results carried out attacks on port 80 using the software DoSHTTP obtained the results that the attack scenario can be processed by NFAT machine with an average success rate of 89,02%.

Results of testing of port 80 scenario attacks using DoSHTTP software can also be presented in graphical form as shown in Fig. 11.

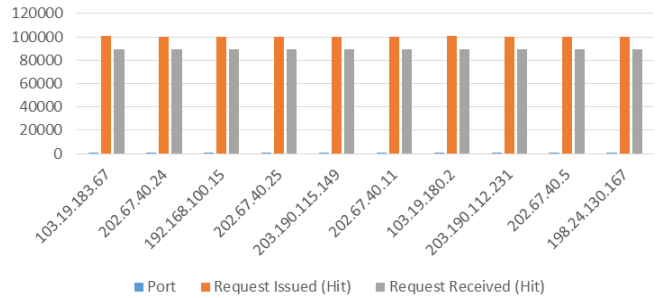


Fig. 11. Graph the results of attack scenario using port 80 DoSHTTP

VI. CONCLUSION

Framework for forensic Internet generated in this study allows users, in this case inverstigators to know the attacks level-related to the attacker's information and resources that is going on in the Internet. This Framework was developed using two stages, namely the clustering process stages and phases of database storage and search logs improved performance. Clustering techniques used in this research was able to classify the level of attacks and shows the attacker information occurs in a network the Internet. Clustering algorithms used machine NFAT (Network Forensic Analysis Tools) using the k-means algorithm. Results of traffic data that is captured in the network is stored in a database for later will be processed using the k-means algorithm to classify the level of attacks into three categories, namely a dangerous attack, rather dangerous attack, and not dangerous attack. The result of testing NFAT machine demonstrate and inform attacks level as well as the information about the attacker that happen in the Internet network with 89,02% success rate to ease the verification process of the source of the attack.

NFAT machine can eventually expanded to be able to detect all the protocols that are commonly used in communications networks, especially the Internet. Additionally, NFAT machine can be installed in some portable devices or embedded so that it has smaller size dimensions, light weight and efficient in the use of power.

ACKNOWLEDGMENT

The authors would like to thank Ahmad Dahlan University (http://www.uad.ac.id) to the research funding.

REFERENCES

[1] Jingna.L, An Analysis on DoS Attack and Defense Technology, The 7th International Conference on Computer Science & Education (ICCSE) July 14-17, Melbourne, Australia, 2012.

[2] Anstee.D, *Worldwide Infrastructure Security Report*, vol. 7," Arbor Networks, Feb. 2012, www.arbornetworks.com/report

[3] NIST-a, Information Testing Laboratory, *Computer Forensics Tool Testing Program*, 2012, www.cftt.nist.gov

[4] NIST-b, *Guide to Integrating Forensic Techniques into Incident Response*, 2012, http://csrc.nist.gov/publications/nist-pubs/800-86/SP800-86.pdf

[5] Hunt R. *New Developments In Network Forensics*–Tools and Techniques, Proceedings of the IEEE, 2012, pp. 376-381.

[6] Pilli, *A Generic Framework for Network Forensics*, International Journal of Computer Applications (0975-8887), 2012, vol. 1, pp. 1-6

[7] Palmer, G, A Road Map for Digital Forensic Research, 1st Digital Forensic Research Workshop, New York, 2001, pp.15-30.

[8] Riadi.I, Istiyanto.J.E, Ashari.A, Subanar, Log Analysis Techniques using Clustering in Network Forensics, *International Journal of Computer Science and Information Security (IJCSIS)*, 2012, vol. 10, pp. 23-30.

[9] Liao.SH., Chu.P.H., Hsiao.PY., *Data Mining Techniques and Application - A Decade review from 2000 to 2011*. Expert Systems with Application, 2012, pp. 11303-11311.

[10] Abbas.O.A, *Comparisons Between Data Clustering Algorithms*, The International Arab Journal of Information Technology, 2008, vol 5, pp. 320-325.

[11] Casey, E. *Handbook of computer crime investigation: forensic tools and technology*. 2004, Academic Press

[12] Mabuto. E.K, H. S Venter2, *State of the art of Digital Forensic Techniques*, proceeding of Information Security South Africa Conference, Department of Computer Science, University of Pretoria, Pretoria, 2011.

[13] Aichi.H, A.Hellany & M.Nagrial, Network Security Approach for Digital Forensic Analysis, 2008.

[14] Strauss. T, Martin S. Olivier, *Network Forensics in a Clean-Slate Internet Architecture*, Proceedings of the IEEE, 2011.

[15] Sridhar N, Dr.D.Lalitha Bhaskari, Dr.P.S.Avadhani, *Plethora of Cyber Forensics*, (IJACSA) International Journal of Advanced Computer Science and Applications, 2011, vol. 2, pp. 110-114

[16] Liu. J, Guiyan.T, *Design and Implementation of Network Forensic System Based on Intrusion Detection Analysis*, International Conference on Control Engineering and Communication Technology, 2012.

[17] Raftopoulos E, Matthias E, and Xenofontas D, *Shedding Light on Log Correlation in Network Forensics Analysis*, 2012.

[18] Chennaka. A, *Network Forensics : A Survey*, Electrical and Computer Engineering, Iowa State University, 2013.

[19] Chen J C, Jiang M C Liu, *Wireless LAN security and IEEE 802.11i*, IEEE Wireless Communications, 2005, pp. 27-36

[20] Xing. X.Y, Shakshukie B., *Security analysis and authentication improvement for IEEE 802.11i specification*, Proc of IEEE GLOBECOM, 2008, pp. 1-5.

[21] Sheng. Y. Tank Chen G, *Detecting 802.11 MAC layer spoofing using received signal strength* C Proc of IEEE, 2008, pp. 1768-1776.

[22] Bagus A., Ali S, Ardelia H., *The design of a mazesolving system for a micromouse by using a potential value algorithm*, Journal World Transactions on Engineering and Technology Education, 2006, pp. 509-512

[23] Han J. and Kamber M., *Data Mining : Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[24] Kumaravel. A, *Multi-Classification Approach for Detecting Network Attacks*, Proceedings IEEE Conference on Information and Communication Technologies, 2013.

[25] Ghosh.S., Sanjay Kumar Dubey, *Comparative Analysis of K-Means and Fuzzy C-Means Algorithms*, (IJACSA) International Journal of Advanced Computer Science and Applications, 2013, vol. 4, pp. 35-39

[26] Mukkamala, S. and Sung, A.H. *Identifying significant features for network forensic analysis using artificial techniques*. International Journal of Digital Evidence, 2003, vol. 1, pp. 1-17

[27] Petersen, J.P. *Forensic examination of log files*. MSc thesis, Informatics and Mathematical Modelling, Technical University of Denmark, Denmark, 2005.

[28] Jacobson, Van, Craig Leres, and Steven McCanne, tcpdump - dump traffic on a network, UNIX man pages, 1998.

[29] Fauziah L, *Computer Network Attack Detection Based on Snort IDS with K-means Clustering Algorithm*, ITS Library, 2009.

[30] Haris,S.H.C, Shadoon, M.G, Ahmag, Ghani, *Anomaly Detection of IP Header Threats*, International Journal of Computer Science and Security (IJCSS), 2011, vol. 4, pp. 497-504.

Consumer Acceptance of Location Based Services in the Retail Environment

Iris Uitz
Graz University of Technology
Institute of Business Economics and Industrial Sociology
Kopernikusgasse 24/II, 8010 Graz,
Austria

Roxane Koitz
Graz University of Technology
Kopernikusgasse 24/II, 8010 Graz,
Austria

***Abstract*—Smartphones have become a commodity item. In combination with their seemingly infinite extensions through mobile applications, they hold great economic potential for businesses. Location Based Services (LBS) take advantage of their portability by providing relevant information to the user regarding their location. Utilizing the user's position to create personalized location-specific marketing messages enables businesses to yield value for their customers. The main objective of this paper is to identify factors, which influence the acceptance of LBS apps in the context of retail since there is a lack of research in this field. A qualitative research approach was chosen to investigate the relevant variables. Based on the conducted interviews, theories were derived and verified against further data retrievals. Similar to findings of previous research the factors ease of use (usability) and usefulness were confirmed as being crucial in forming consumers' attitudes.**

Keywords—Location Based Services; Usability-Testing; Apps; Shopping-Apps; Mobile; Consumer Acceptance; Mobile Marketing

I. Introduction

Location Based Services (LBS) utilize a device's position in order to provide the consumer with personalized information. A widespread application of LBS is navigation systems. There are several methods on how to achieve the positioning of a mobile device. A common technique is to use satellite-positioning systems like the Global Positioning System (GPS). However, this method is not feasible for indoor navigation due to weakened signal strengths and inaccuracy. For the positioning indoors, two different approaches are generally applied. The first method uses a signal infrastructure consisting of Wi-Fi signals, for examples. The determination of the device's position can be calculated based upon signal strengths. The alternative is a dead reckoning approach using an inertial system. Those systems take a starting point and calculate the position based upon sensor measurements of acceleration and possibly other variables.

LBS are by far not limited to navigation; they provide the capabilities to be used for a large amount of purposes. Especially the creation of a service allowing an accurate positioning inside combined with smartphones poses unique opportunities for businesses. Owing to the portability of smartphones and their ubiquitous Internet access retailers can use them to market their products virtually whenever and wherever. By taking the user's location information into account, companies can create context specific offers and notifications. They hence ameliorate consumer targeting. Location awareness is an especially interesting aspect, since more than half of the purchase decisions are made within the stores themselves [14]. In regard to this finding, location-based marketing strategies, addressing the consumer at the point of the purchase decision, do not only hold potential for the entrepreneur, but also create value for the consumer through relevant personalized messages. There are already several mobile applications integrating LBS in retail environments. Ma$$iv€, as described in [2], for instance, is an intelligent mobile grocery shopping assistant app and web application. The system is based on the user's mobile shopping list and is able to navigate the consumer to items on the list within the store. Furthermore, it provides the user with product and recipe suggestions as well as offer notifications. Another mobile shopping application is aisle411[1], which enables the consumer to create shopping lists and displays the item's location within the branch. In addition, personalized offers and advertisements can be integrated within the app. As for right now, an LBS app for a Swiss retailer is under way, which should also have similar capabilities. In addition to an indoor navigation, the app will provide context and location aware offers and notifications.

The central question of this paper is to investigate factors, which influence the consumer acceptance of LBS in the context of retail to unfold its potential. There is a vast amount of studies and models on variables affecting the consumer's intention. Some of the most prominent models include the Technology Acceptance Model (TAM), which identifies the perceived usefulness and the perceived ease of use of a system as factors that have a direct effect on the attitude a consumer forms towards it [5]. This attitude eventually determines the actual use. Reference [3] expanded the TAM for handheld Internet devices and appended the perceived enjoyment to the drivers of the consumer's attitude. Specializing on mobile shopping services [13] confirmed the relevance of usefulness and ease of use in this context. Despite the quantity of research available, the special case of LBS in the retail environment has been neglected so far. On grounds of future mobile applications, such as the Swiss retailers, this use case is worth investigating.

II. Theory

LBS delineate services, which take the users' location into account in order to provide them with relevant information. Generally, two types of LBS can be distinguished [8]. The first

[1] http://aisle411.com/

type sends location-based information to the device after a prior user request. These kind of services are called Pull-based, whereas their counterpart and second type are Push-based services, which are not triggered by a direct user request. Another possibility to classify LBS is in accordance with their application [6]:

Information/Directory Services provide information on nearby points of interest as; for example, hospitals or restaurants.

Tracking/Navigation Services comprise, for example, navigation systems or directions.

Emergency Services are for instance roadside assistance or police and firemen response.

Location Based Advertising implies activities like mobile coupons, personalized advertising or offers.

A. *Mobile Marketing*

Resulting from the pervasiveness of smartphones, the relevance of those devices as far as marketing is concerned is constantly increasing. According to the Mobile Marketing Association (MMA) mobile marketing consists of all activities companies perform in order to use mobile devices and networks as a channel to communicate and interact with consumers [12]. The advantage of mobile marketing lies in the ubiquity of smartphones, which facilitates a virtually constant communication with the user.

Furthermore, the portability of smartphones allows expanding context specific advertising by considering location-based information of the device; hence creating more personalized marketing messages [6]. This subcategory of mobile marketing is called Location Based Marketing and the advertising activities fall under the term Location Based Advertising (LBA). Analogous to the types of LBS there are Push-based LBA and Pull-based LBA. The MMA gives an overview of several possible marketing strategies, which can be achieved by LBS [12]:

Geo-targeted Text and Display Advertising Placement of advertising messages within mobile media integrating LBS. The promotion message can either be sent to users in a certain geographic area (User Targeting) or to the user's device depending on their location (Message Targeting).

Embedded Icons The personalized advertising is embedded within maps, apps, or on web pages. Icons, or logos of the sponsors are displayed on the map to show the user the relative distance between their location and the sponsor's location.

Search (aka Local Directory Advertising) Used by directory services like the Yellow Pages and provide listings of local companies in regard to proximity to the user's location.

Location Triggered Notifications After an opt-in those services provide advertising messages to the consumer when in a certain range to a merchant or whenever special offers are available.

Location Branded Applications Describe mobile applications, which have a LBS integrated.

Check-in Based Contests and Games User's can "check-in" using the LBS portion of the mobile media to earn rewards such as discounts or mobile coupons depending on their location.

Click-to-X Routing Calls for broad campaigns are being routed to local call centers.

As earlier mentioned, more than half of the purchase decisions are made within the store [14]. Therefore marketing strategies, which focus on the consumers at the act of shopping itself promise to be fruitful. In this context purchases can be divided for example, along the axis of the amount of planning involved anterior to the purchase in spontaneous and deliberate purchases (see Table I) [1]. LBA messages can trigger different purchase behaviors. For instance, mobile coupons delivered depending on the user's position can trigger promotional purchase behavior. However, not only reward based LBA strategies prove to yield results. LBA works with as well as without incentives [16]. Notifying the consumer about a product as they pass by it can possibly cause impulsive purchases. Hence, there are several possibilities in which purchases can be facilitated by considering the consumer's location.

B. *Consumer Acceptance*

In the literature there is a great amount of research on the drivers of behavioral intention, whether focusing on technology, or in general. Many models and studies are based upon results of previous investigations; hence being similar to one another and identifying comparable concepts as relevant to consumer acceptance.

TABLE I. PURCHASE BEHAVIOR [1]

Deliberate Purchases	Extended Purchase Decision Making	Making a purchase based on objective, logical criteria and for utilitarian reasons.
	Symbolic Purchase Behavior	Buying a brand to project a certain image or because it meets with social approval.
	Repetitive Purchase Behavior	Making a routine purchase or buying something because you're loyal to it.
	Hedonic Purchase Behavior	Buying something because you just like it.
Spontaneous Purchases	Promotional Purchase Behavior	Buying something because it's on sale.
	Exploratory Purchase Behavior	Buying something out of curiosity or because of a desire for variety.
	Casual Purchase Behavior	Buying something without thinking much about it.
	Impulsive Purchase Behavior	Buying something on impulse.

Consumer Technology Acceptance Model The Consumer Technology Acceptance Model (c-TAM) is based on the TAM described and emphasizes the consumer context. The TAM as well as its expansions serve the purpose of trying to explain how consumer acceptance of a technological system in a certain context forms. The model proposed in [3] focuses on handheld Internet devices. The initial factors influencing consumer acceptance are comprised under the term "External Variables" (see Fig. 1). These take the consumer's visual

orientation as well as the type of device into consideration. Users tending to process information visually find it easier to perform tasks on Internet devices, which is the reason of the causal relation between the consumer's visual orientation and the Ease of Use (EOU). The device itself as well as its characteristics influences the EOU as well, since a bigger screen; for example, facilitates usage. Furthermore, the variable FUN, describing the expected enjoyment of using a system, can vary from device to device. Reference [3] elucidate that handheld devices boost the consumer's intrinsic motivation and hence their enjoyment. In addition, there is a direct influence from the EOU on FUN. This relation results from the notion that the system use is more enjoyable if the user can perform tasks without hindrance emerging from the system itself. On the cognitive level, EOU further influences the Perceived Usefulness (USEF). Usefulness and FUN determine the consumer's attitude towards utilizing the system, which further influences the Behavioral Intention. There have been several studies conducted aiming at determining the prediction of user behavior. Reference [13] for example, focused on mobile services in their research. According to the study, issues concerning the screen size and therefore worse readability have become irrelevant to a certain point. However, obstacles such as economic factors preventing a widespread reachability through mobile services remain. Following TAM it can be shown that USEF, as well as EOU positively influence the acceptance of mobile phones in the retail environment [13].

Quality of Service Three key factors were established in having an influence on consumer acceptance of mobile Internet services: service quality, satisfaction, and value [17]. Upon mobile applications, Quality of Service was identified as a criterion for app acceptance [18]. The Quality of Experience can be partially ascribed to the design. Research results point to the assumption that especially the interfaces as well as the interaction with the app play a role in the perception of app design [10].

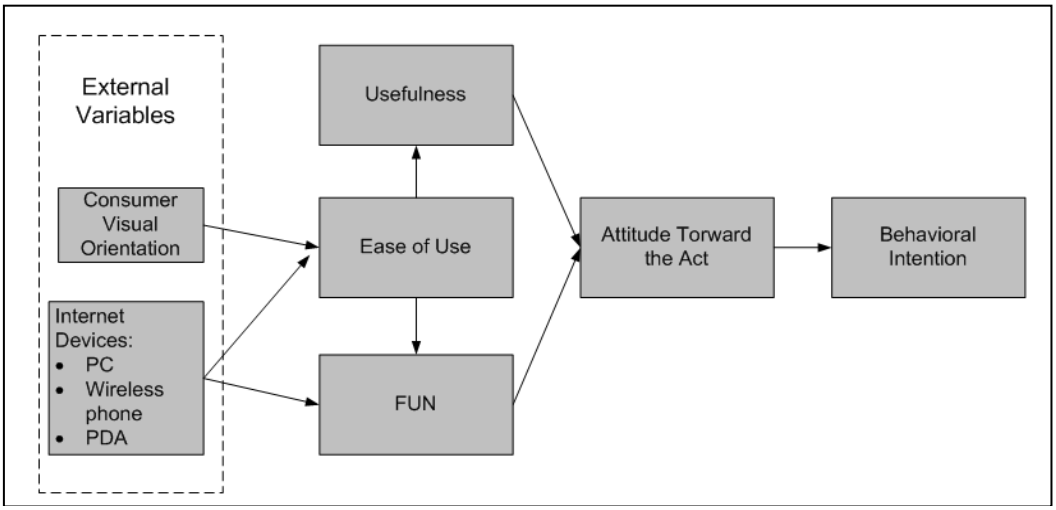


Fig. 1. c-TAM

Utilitarian/Hedonic Performance Expectancy The consumer's attitude towards services; hence, the behavioral intention to use those services, is determined by two expectations. On the one hand, the Utilitarian Performance Expectancy in this context includes variables like flexibility or effectiveness of the services. On the other hand, there are Hedonic Performance Expectancies, which consider the enjoyment, the user experiences, while using the service. Furthermore, Effort Expectancy, associated with the work necessary to use the service, influences Utilitarian as well as Hedonic Performance Expectancy. Besides, those Social Influences and Facilitating Conditions affect the intention [20].

Motivation Theory Reference [9] investigated the acceptance of LBS in regard to the motivation theory. It was shown that intrinsic motivation, which subsumes the enjoyment of usage, influences the consumer's intention to a greater extent than extrinsic motivation. Pressure or incentives are attributed to extrinsic motivation. In addition it was shown that the initial extrinsic motivation does not allow drawing conclusions with respect to further application.

Value The value is described as the balance between benefits and risks. The risks in the context of LBS lay within the sharing of the user's position. The benefits are, for example, personalized offers [16]. The personalization of offers was found to positively influence the consumer's attitude [11]. Furthermore, entertainment has a positive influence on attitude.

Critical Success Factors In [4] a study was conducted identifying several critical success factors for LBS. In the first portion of the study, the participants were to name factors important regarding LBS. The criterion named most often by the test subjects within the brainstorming sessions was privacy. Privacy concerns about LBS are one of the most prominent factors, which have a negative influence on the consumer acceptance. In the context of privacy Personalization Privacy Paradox is described as following [19]: As personalization improves, concerns in regard to privacy increase. Hence, the consumer acceptance decreases. Reference [14] identified the disclosure of the user's home as the greatest concern. The second portion of the study was designed to rank the most frequently named factors of the first portion in a quantitative

survey. The ranking of the fifteen examined factors are displayed in Table II.

TABLE II. CRITICAL SUCCESS FACTORS AND RANKING [4]

Critical Success Factors	
1. Speed	9. Smart Location-Based Services
2. Real-Time/Up-To-Date Information	10. Aesthetics
3. Cost	11. Quality of Reviews
4. Usefulness	12. Integrated Applications/Services
5. Simple/Ease of Use	13. Standards & Platform Independence
6. Reliability	14 .Size of Application
7. Personalization/ Preference Setting	15. Publicity of Location-Based Services
8. Privacy	

The discrepancies within the first and second part might be resulting from the focus on different age groups. Most of the participants of the brainstorming sessions were between 36 and 40 years old and had achieved a doctor's degree. In contrast to the first portion of the research, where mostly test subjects between the ages of 15 and 25 years participated.

III. METHOD

A. Grounded Theory

The qualitative research approach Grounded Theory was chosen so as to investigate the research question. A study conducted in accordance with the Grounded Theory origins from an initial data collection from sources varying in their characteristics. Once the initial data collection is completed, codes for the data are developed. Coding is used to find significant portions of information within the data and connecting those into concepts and categories. After the extraction of the concepts and categories, those are used to write memos and derive hypotheses relevant to the research area. Those hypotheses, however, have to be verified. Therefore, new data are collected and coded. Furthermore, it has to be determined to which extent the theories can be confirmed by the new data. In accordance with [7], if discrepancies occur between the data and the theories the theories have to be adjusted. The cycle of collecting and coding data, developing theories, and verifying those based on new data is called Theoretical Sampling and is repeated until theoretically there is no more information gain possible. This state is known as the Theoretical Saturation. A schematic overview of the Grounded Theory is shown in Fig. 2.

B. Data Collection

The data collection method used within this study was a General Interview Guide Approach as described in [15]. The interview guide was revised at the end of each Theoretical Sampling Iteration in order to verify the developed theories. Fifteen participants were interviewed. The interviews consisted of four parts:

Background Information The first portion of the conversation was used to collect demographic data.

Smartphone Usage The second part was concerned with the participant's experience with smartphones. For example, they were asked to name criteria of good mobile applications and what has been bothering them about apps they have already used (e.g. *"Which factors compose a 'good' app in your opinion?").*

Buyer Behavior The third part consisted of questions concerning their buying behavior. For instance, the test subjects were asked whether their buying behavior had changed over the past years and if so, which environmental factors had caused those changes (e.g. *"Would you say that your buying behavior has changed within the last couple of years? If so, which factors have been responsible for those changes?").*

Shopping Apps The last portion of the interview covered the participant's opinion about mobile applications in the retail environment, focusing on LBS in more detail. Therefore, several scenarios were presented to the interview partners. In those scenarios the participants were grocery shopping and used a smartphone shopping app, which allowed them; for example, to create a mobile shopping list, showed them different kinds of offers or navigated them to products found on their shopping list. After each scenario the subjects where asked about different parts of the scenario and their receptive perception (e.g. "*Imagine the following scenario: You are grocery shopping in your usual supermarket to buy the following items: Milk, Tea, Baking Powder, Canned Corn, Chocolate. You probably know where those products are located within your supermarket. Furthermore, imagine that you do not have a handwritten shopping list, but an app, which contains a mobile version of the shopping list. Would you like to receive offers of this branch on your smartphone or would this bother you? Would you like to receive offers of the items on your shopping list on your smartphone or would this bother you? Would you be interested in offers of items on your shopping list but from different brands?").*

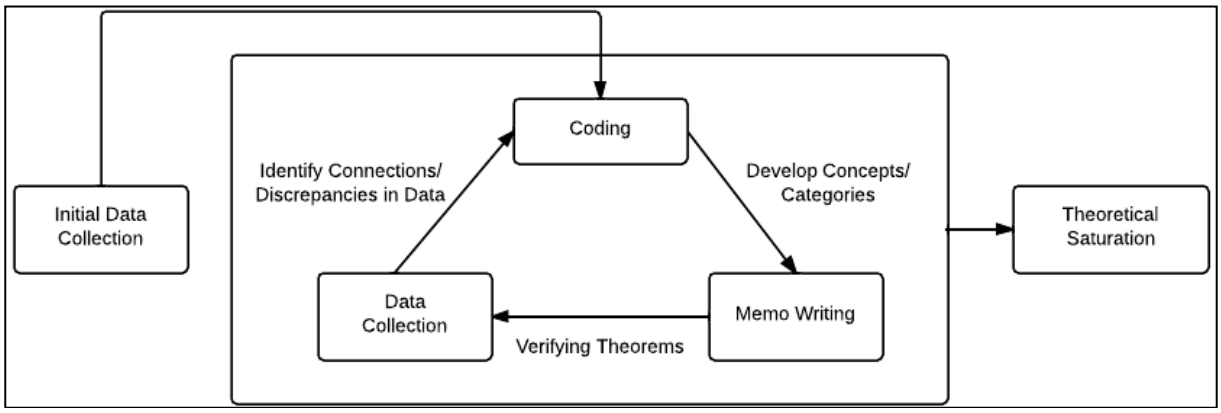


Fig. 2. Grounded Theory schema

Towards the end, the participants were asked to rank eight different factors in accordance with their importance concerning a shopping app as described in the scenarios. The most important factor should be ranked first and the least crucial one last. ("*Within the*

context of the scenarios, please rank the following factors according to their importance for a shopping app and please explain your ranking: Usability, Cost, Reliability, Speed, Design, Usefulness, Battery Consumption, Up-To-Date Information.").

C. Research Process and Hypotheses

Fig. 3 gives an overview of the research process. Throughout the study different hypotheses were developed from the insights obtained out of the data collection. In the process of verifying these resulting from a larger amount of information from the newly acquired data some of the hypotheses have been refined throughout the study. Furthermore, as a result of the new findings, the interview guide was adjusted at the end of each iteration to accommodate for those modifications. All in all twenty-one theorems were laid down. The ten most prominent are mentioned in Table III.

TABLE III. HYPOTHESES

H2'	*The buyer behavior changed within the last years for most consumers; however, no clear tendencies can be observed.*
H3	*Shopping apps in the retail environment are not widespread; consumers prefer traditional shopping lists to mobile shopping lists.*
H6'	*Consumers have a positive perception of offers of the branch as well as for items on their shopping list.*
H7	*Offers concerning items of the shopping list but from other brands are accepted by most of the consumers*
H8'	*Consumers feel ambivalent about offers customized to the consumer's purchase behavior.*
H9	*Privacy concerns in regard to the app do not determine the consumer acceptance.*
H12'	*Navigation is perceived as helpful within unknown branches and under time pressure. However, some of the consumers determine the relevance of the navigation regarding the size of the branch.*
H15	*Usability is the most crucial factor influencing the consumer's perception of a 'good' app.*
H17	*Apps have to be reliable; otherwise, the consumer is frustrated.*
H21'	*The more an app fulfills the crucial factors; the more the consumers are willing to cut back on less important criteria.*

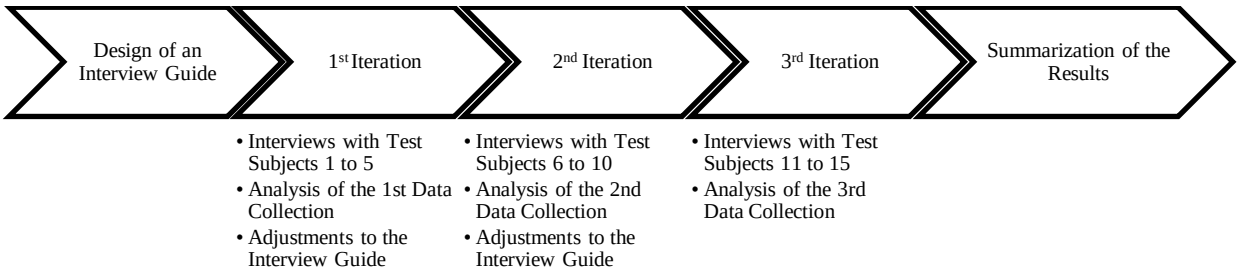


Fig. 3. Research process

IV. RESULTS

A. Participants

Three iterations of the Theoretical Sampling cycle were conducted. Within each iteration five participants were interviewed; this corresponds to a total of 15 test subjects (7 female, 8 male). Further 53.3% of the participants had obtained some kind of degree within a technological field. As shown in Fig. 4, most of the participants were between 20 to 29 years old at the time of the study. The focus on this age group is, on the one hand, due to the greater adoption of technology within younger age groups. On the other hand, consumers younger than this usually have no household of their own and therefore do not necessarily have a distinct buying behavior in regard to groceries on which this study focused.

B. Hypotheses

All the hypotheses presented in this paper could be confirmed. However, this is resulting from the adjustments of the theories necessary through the review of newly collected data. Refined hypotheses are marked with an apostrophe (e.g. H2').

Buyer Behavior (H2') All except three test subjects confirmed that their buying behavior has changed within the last couple of years. However, within this study there was no general tendency regarding the way the behavior has changed. The answer mostly given was that the participants purchase different items than before, for example, due to a changed life style, or more cost-effective items based on budgeting concerns. Two out of the three test participants which did not observe a change in their buying behavior were living in a household where they were neither responsible for the grocery shopping itself nor for the financial planning of the expenses.

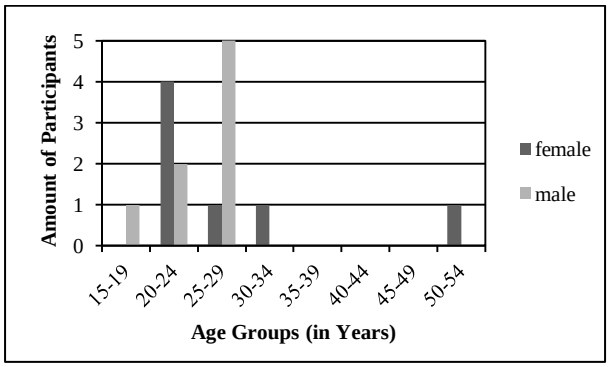


Fig. 4. Overview of participants

Shopping-Apps (H3) None of the interviewed subjects used any of the available mobile apps local retailers provide. The reasons named were that either the participants were unfamiliar with the existence of such apps, or they perceived the available apps as useless. Throughout this section participant commentaries are italicized in this portion of the paper and were translated from the interview transcripts.

"Well, in regard to mobile shopping lists, because for me they are too cumbersome and I am faster using a piece of paper and writing down my products than opening the app and typing them in using the keyboard on the smartphone … And, yeah, I have not found a benefit in the apps of the retailers. But, I have to admit that I have not tried them out yet."

Offers (H6', H7, H8') In regard to the offers made available through a shopping app the perception of the consumers depend on the kind of offer made. An overview of the acceptance of different kinds of offers is depicted in Fig. 5. Eight of the participants would be interested in receiving offers of the branch in which they are shopping and twelve would like to receive offers which are customized to their shopping list (H6').

In regard to items of the shopping list from other brands more than half of the participants would be interested. However, some of the subjects pointed out that the item itself plays a role in whether an offer would be accepted or not. For instance, dairy products are one of the food categories, which created a greater brand loyalty than others within this study (H7).

"… it would be important to me, I don't know, if I for example think about dairy products or eggs, there I, it is important to me that it is a renowned brand. However, I would not care about chocolate."

All the test subjects confirmed that offers of the products regularly purchased are useful and are perceived in a positive way. This, however, is not true for a certain scenario in which the participants were presented with the offer of an item, regularly purchased by them although not on the shopping list, as they pass by this specific item in the grocery store. Eight of the participants perceived this as an invasion of privacy or just generally negative. Although not all the test subjects were interested in every kind of offer, the negative perception of those offers was low among the subjects except for this one.

"The emotion in regard to this is ambivalent; of course, it would be practical to be reminded, since the supply of different groceries is coming to its end, on the other hand, this [reminder] would be a paternalism, which would probably start being unpleasant."

However, the option of setting preferences in receiving offers was widely accepted and found useful by the questioned participants (H8').

Privacy (H9) Twelve of the subjects had no serious concerns in regard to their privacy being invaded by the shopping app presented in the scenarios in general. Except for three participants, who would decide if they would use the app after reading the terms of use.

"As for right now my concerns would be minor, since I don't believe that any tenuous information could be derived…"

Navigation (H12') The navigation within a grocery store was perceived as helpful by all but two participants. Some of the test subjects, however, can only imagine using it within unknown stores, when they are under time pressure or stores with a great sales area.

"Well, that would be helpful. Very often I have been shopping in stores, and I could not find certain items and had to find employees to ask them where this product is and the employees did not know exactly where to find the item either and then this could really facilitate this if it would really work and tell me at least in which aisle it should be located or show me the way to the location. However, it would be very bothering if the item was not at that location…"

Factors (H15, H17, H21') Different factors influencing consumer acceptance were identified before as well as during the study. At the beginning of the interviews all fifteen participants were asked to name factors, which according to their opinion characterize a good app. The amount of participants naming a specific factor is shown in the last column of Table IV. Eight factors were identified through the most prominent results of the first iteration as well as drivers found in the literature.

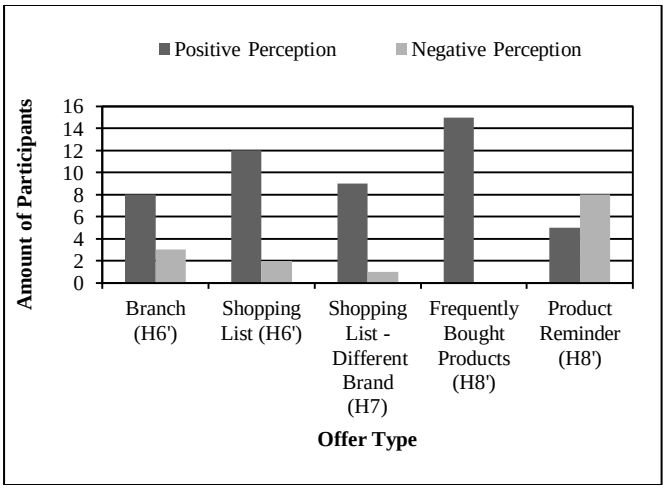


Fig. 5. Consumer acceptance of offers

The test subjects of the second and third iteration were asked to rank those eight factors in regard to their importance for a shopping app as described in the scenarios presented to them before. Those answers determined the ranking as well as the rank of the factors available in the first and third column of Table IV.

Usability was the highest ranked factor as well as named by all except one participant. Besides this clear ranking, when asked, which factors are somewhat interchangeable, Usability was named by several participants as not interchangeable. Reliability and Usefulness reached the same ranking. However, Usefulness was named fewer times by the participants.

"Another reason, when apps crash then I normally just leave them and not open them a second time"
"Usability is also important to me, I believe that it is not enjoyable when you use an app and have no idea on how to use it."

TABLE IV. EVALUATION OF FACTORS INFLUENCING ACCEPTANCE

Rank	Factor	Ø Ranking	Times Mentioned
1	Usability	2,6	14
2	Reliability	2,9	6
2	Usefulness	2,9	1
4	Cost	3,0	5
5	Speed	3,3	9
6	Up-To-Date Information	3,7	0
7	Design	5,1	8
8	Battery Consumption	5,6	1

V. DISCUSSION AND CONCLUSION

The general notion from this study is that consumers within the investigated age group would accept a LBS application in the context of retail. Throughout the literature on technology acceptance in general, on LBS or on mobile services a few factors are similar amongst the models and theories. Taking the c-TAM as a starting point, the three crucial factors influencing the consumers' attitude hence determining their intention to utilize a system are usefulness, ease of use, and fun. The usefulness of a system was identified as a critical variable in many studies, although the concept was being delivered under different terms. Regarding the motivational theory, for example, some of the characteristics of usefulness can be ascribed to extrinsic motivation. Utilitarian Performance Expectancy is another similar variable determining attitudes towards technology. Besides, regarding the results found within the literature, the conducted study confirmed the importance of usefulness within the context of LBS in retail as the second most prominent factor identified. The factor of reliability is also accounted to usefulness.

Reference [3] included the ease of use within their model. The concept was processed within the study under the term of the usability of the proposed application. Besides, the extensive research done in the field of the EOU as a critical factor in influencing behavioral intention in different fields of expertise [13] as well as the study conducted confirmed its eligibility for mobile shopping services.

Comparing the results from this study to [4] it is clear to see that even though the ranking of the factors is not exactly equivalent, the six most crucial factors are the same in both studies as can be seen in Table V.

The factor of fun or enjoyment was not investigated within this study. The only mention of this within the conducted study focused on the effect reliability as a software quality characteristic has on the user enjoyment. However, different literature sources suggest that enjoyment, intrinsic motivation, or hedonic performance expectancy does play a role.

TABLE V. FACTOR RANKING OF THE STUDY IN COMPARISON TO [4]

Factor	Rank Study	Rank [4]
Usability (Ease of Use)	1	5
Reliability	2	6
Usefulness	2	4
Cost	4	3
Speed	5	1
Up-To-Date Information	6	2
Design	7	10
Battery Consumption	8	-

In the literature, the concepts vary among the authors in the sense that the terms describe similar drivers, variables of one theory merge within different factors of others and some characteristics of the concepts are disregarded in some studies whereas they are the focal point of others. However, the general idea of the dominant factors from the point of the study conducted as well as the review of the literature is best described by the c-TAM as well as its predecessors, which were not mentioned in detail within this paper.

The factor privacy was not a determining factor of usage within this study. This might be due to two reasons. First, the app described within the scenarios is an example of a pull-based LBA strategy. Those tend to create fewer privacy concerns than their counterparts. The second reason might be ascribed to the fact that the determining of the location was solely conducted within the store. Reference [14] showed that the greatest privacy issue for consumers is the disclosure of their home, this use case has no relevance within the study.

Regarding the scenarios delivered to the participants, the most prominent finding was that the acceptance of personalized offers depends on the context in which the shopper is as well as on the way the offer is delivered. Hence the point where the positive perception of a buying suggestion shifts towards a negative notion is worth investigating further.

With regard to future research, other factors known from the literature like fun and social influences should be investigated in the context of location-aware apps in retail. Further research on the findings of this study with different age groups would provide a more general picture of the research field.

REFERENCES

[1] Baumgartner, H., "Toward a Personology of the Customer", Journal of Consumer Research, vol. 29, no. 2, 2002, pp. 286-292.

[2] Bhattacharya, S., Floréen, P., Forsblom, A., Hemminki, S., Myllymäki, P., Nurmi, P., Pulkkinen, T., and Salovaara, A., "Ma$$iv€ - An Intelligent Mobile Grocery Assistant", 8th International Conference on Intelligent Environments Guanajuato, Kune, 2012, pp. 165.

[3] Bruner II, G.C., and Kumar, A., "Explaining consumer acceptance of handheld Internet devices", Journal of Business Research, vol. 58, no. 5, 2005, pp. 553-558.

[4] Chin, N., "Critical Success Factors of Location-Based Services", Dissertation, University of Nebraska-Lincoln, 2012.

[5] Davis, F.D., "A technology acceptance model for empirically testing new end-user information systems: theory and results", Dissertation, Massachusetts Institute of Technology, Sloan School of Management, 1985.

[6] Dhar, S., and Varshney, U., "Challenges and business models for mobile location-based services and advertising", Commun.ACM, vol. 54, no. 5, 2011, pp. 121-128.

[7] Flick, U., von Kardorff, E., and Steinke, I., Qualitative Forschung - Ein Handbuch, 7th edn, Rowohlt Taschenbuch Verlag, Hamburg, 2009.

[8] Hilty, L., Oertel, B., Wölk, M., and Pärli, K., Lokalisiert und identifiziert. Wie Ortungstechnologien unser Leben verändern., 1st edn, Vdf Hochschulverlag, Zurich, 2012.

[9] Ho, S.Y., "The effects of location personalization on individuals' intention to use mobile services", Decision Support Systems, 2012.

[10] Ickin, S., Wac, K., Fiedler, M., Janowski, L., Jin-Hyuk Hong, and Dey, A.K., "Factors influencing quality of experience of commonly used mobile applications", Communications Magazine, IEEE, vol. 50, no. 4, 2012, pp. 48-56.

[11] Lee, Y., "Factors influencing attitudes towards mobile location-based advertising", Software Engineering and Service Sciences (ICSESS), 2010 IEEE International Conference on, 2010, pp. 709.

[12] Mobile Marketing Association, "Mobile Location Based Services - Marketing Whitepaper", Mobile Marketing Association, 2011.

[13] Shankar, V., Inman, J.J., Mantrala, M., Kelley, E., and Rizley, R., "Innovations in Shopper Marketing: Current Insights and Future Research Issues", Journal of Retailing, vol. 87, Supplement 1, 2011, pp. 29-42.

[14] Tsai, J., Kelley, P.G., Cranor, L.F., and Sadeh, N., "Location-Sharing Technologies: Privacy Risks and Controls", TPRC, 2009.

[15] Turner, D.W., "Qualitative Interview Design: A Practical Guide for Novice Investigators", The Qualitative Report, vol. 15, no. 3, 2010, pp. 754-760.

[16] Unni, R., and Harmon, R., "Perceived effectiveness of push vs. pull mobile location-based advertising", Journal of Interactive Advertising, vol. 7, 2007, pp. 1-24.

[17] Vlachos, P.A., and Vrechopoulos, A.P., "Determinants of behavioral intentions in the mobile internet services market", Journal of Services Marketing, vol. 22, no. 4, 2008, pp. 280-291.

[18] Wac, K., Ickin, S., Hong, J., Janowski, L., Fiedler, M., and Dey, A.K., "Studying the experience of mobile applications used in different contexts of daily life", Proceedings of the first ACM SIGCOMM workshop on Measurements up the stack ACM, New York, NY, USA, 2011, pp. 7.

[19] Xu, H., Luo, X., Carroll, J.M., and Rosson, M.B., "The personalization privacy paradox: An exploratory study of decision making process for location-aware marketing", Decision Support Systems, vol. 51, no. 1, 2011, pp. 42-52.

[20] Yang, K., "Determinants of US Consumer Mobile Shopping Services Adoption: Implications for Designing Mobile Shopping Services", Journal of Consumer Marketing, vol. 27, no. 3, 2010, pp. 262- 270.

Recognition of Objects by Using Genetic Programming

Nerses Safaryan
Algorithmic Languages and Programming,
State Engineering University of Armenia
Yerevan, Armenia

Hakob Sarukhanyan
Digital signal and Image processing laboratory,
Institute for Informatics and Automation Problems of
National Academy of Sciences
Yerevan, Armenia

***Abstract*—This document is devoted to the task of object detection and recognition in digital images by using genetic programming. The goal was to improve and simplify existing approaches. The detection and recognition are achieved by means of extracting the features. A genetic program is used to extract and classify features of objects. Simple features and primitive operators are processed in genetic programming operations. We are trying to detect and to recognize objects in SAR images. Due to the new approach described in this article, five and seven types of objects were recognized with good recognition results.**

Keywords—Terminals; Fitness; Selection; Crossover; Mutation; Ground Truth

I. INTRODUCTION

Object recognition is still a challenge for computer vision systems in general. The main purpose of object recognition is to identify the kinds of the objects in an image [1]. Object recognition algorithms rely on matching or learning algorithms using appearance-based or feature-based techniques. We are trying to achieve good recognition results using the feature-based technique [2, 6]. The quality of object recognition is heavily dependent on the effectiveness of features. The features used to represent an object are the key to the object detection and recognition. It is difficult to extract good features from real images due to various factors, including noise. There are many features that can be extracted. It is very difficult to find appropriate features and to synthesize composite features. Synthesizing effective new features from primitive features is equivalent to finding good points in the feature combination space where each point represents a combination of primitive features. The feature combination space and feature subset space are huge and complicated and it is very difficult to find good points in such vast spaces unless one has an efficient search algorithm [5, 7]. Genetic programming (GP) is used as search algorithm. GP may try many unconventional combinations and in some cases these unconventional combinations yield exceptionally good recognition performance. Also, the inherent parallelism of GP and the speed of computers allows a much larger portion of the search space to be explored. We have used a simple (steady-state) [2] genetic programming algorithm and primitive features to detect and to recognize objects in SAR images. A program system is developed based on this GP algorithm by means of which two experiments were done: trying to recognize five types of objects, trying to recognize seven types of objects.

II. GENETIC PROGRAMMING IN OBJECT RECOGNATION

Individuals of GP are composite operators in task object recognition. The composite operators are represented by binary trees. Internal nodes of binary trees are primitive operators and leaf nods are primitive features [4]. GP uses five major considerations in the task of object detection and recognition:

A. The set of terminals:

The set of terminals include the following images: F_0, F_1, …, F_{15}, where F_0 is the original image, the F_1, F_2 and F_3 are 3x3, 5x5 and 7x7 mean images, the F_4, F_5 and F_6 are 3x3, 5x5 and 7x7 deviation images, the F_7, F_8 and F_9 are 3x3, 5x5 and 7x7 maximum images, the F_{10}, F_{11} and F_{12} are 3x3, 5x5 and 7x7 minimum images and the F_{13}, F_{14} and F_{15} are 3x3, 5x5 and 7x7 median images [4].

B. Primitive operators:

The primitive operators are given below in the Table 1 [2, 4].

TABLE I. THE SET OF TERMINALS

No.	Operator	Description
1	$ADD(A, B)$	Add images A and B.
2	$SUB(A, B)$	Subtract image B from A.
3	$MUL(A, B)$	Multiply images A and B.
4	$DIV(A, B)$	Divide image A by image B (If the pixel in B has value 0, the corresponding pixel in the resultant image takes the maximum pixel value in A).
5	$MAX2(A, B)$	The pixel in the resultant image takes the larger pixel value of images A and B.
6	$MIN2(A, B)$	The pixel in the resultant image takes the smaller pixel value of images A and B.
7	$ADDC(A)$	Increase each pixel value by c.
8	$SUBC(A)$	Decrease each pixel value by c.
9	$MULC(A)$	Multiply each pixel value by c.
10	$DIVC(A)$	Divide each pixel value by c.
11	$SQRT(A)$	For each pixel with value v, if $v \geq 0$, change its value to v. Otherwise, to $-\sqrt{-v}$.
12	$LOG(A)$	For each pixel with value v, if $v \geq 0$, change its value to ln(v). Otherwise, to –ln(-v).
13	$MAX(A)$	Replace the pixel value by the maximum pixel value in a 3×3, 5×5 or 7×7 neighborhood.
14	$MIN(A)$	Replace the pixel value by the minimum pixel value in a 3×3, 5×5 or 7×7 neighborhood.
15	$MED(A)$	Replace the pixel value by the median pixel value in a 3×3, 5×5 or 7×7 neighborhood.

16	*MEAN*(*A*)	Replace the pixel value by the average pixel value of a 3×3, 5×5 or 7×7 neighborhood.
17	*STDV*(*A*)	Replace the pixel value by the standard deviation of pixels in a 3×3, 5×5 or 7×7 neighborhood.

C. Fitness value

The fitness value of a composite operator is computed in the following way. Suppose *G* and *G′* are foregrounds in the ground truth image and the resultant image of the composite operator respectively. Let n(X) denote the number of pixels within region X, then Fitness = n (G∩G′) / n (G∪G′) [2, 4].

D. Parameters and termination

We have used the following parameters for GP: Population size - M, the number of generation - N, the crossover rate, the mutation rate and the fitness threshold [3, 10]. The GP stops whenever it finishes the pre-specified number of generations or whenever the best composite operator in the population has fitness value greater than the fitness threshold.

E. Operations of Genetic programming

The search is done by performing selection, crossover and mutation operations in GP [2, 3].

a) Selection: genetic operators in GP are applied to individuals that are probabilistically selected based on fitness. Better individuals are more likely to have more child programs than inferior individuals. We are using tournament selection [3]. In tournament selection a number of individuals are chosen at random. These are compared with each other and the best of them is chosen to be the parent.

b) Crossover: we are using sub tree crossover [3]. Two composite operators are selected on the basis of their fitness values as parents: sub tree crossover randomly selects a crossover point in each parent tree, two subtrees rooted at these two nodes are exchanged between the parents to generate two new composite operators which are called offspring.

c) Mutation: mutation is introduced to randomly change the structure of some individuals. We are using tree type of mutation [2, 3]:

- Randomly select a node of the binary tree representing a composite operator and replace the sub tree rooted at this node, including the node selected, by another randomly generated binary tree.
- Randomly select a node of the binary tree representing a composite operator and replace the primitive operator stored in the node with another primitive operator of the same number of inputs as the replaced one. The replacing primitive operator is selected at random from all the primitive operators with the same number of input as the replaced one.
- Randomly select two sub trees within a composite operator and swap them. Of course, neither of the two sub-trees can be a sub-tree of the other.

III. Used algorithm of Genetic Progrmming

Steady-state genetic programming [2, 4] is used to synthesize composite operators. The first step of the algorithm is to randomly generate the initial population. We will use half-and-half method for generating the initial population [3, 4, 6]. Suppose there are a set of primitive operators O = {O_1, O_2, ..., O_N} and a set of terminals F={F_1, F_2, ..., F_M}. If we randomly generate population P of size M from O and F, we obtain their combinations, which consist of primitive operators and terminals, as members (composite operator) of P. Then each composite operator in P is evaluated and crossover operation is perform. If we show composite operators as a binary tree, they will be like on figure 1 A. We can have two offspring (composite operators) after crossover operation, which are shown in figure 1 B. These two offspring replace two of the worst composite operators in P. This is continued until crossover rate is met. Then mutation is performed, which can replace the primitive operator stored in the node with another primitive operator. For example, if we observe the tree of figure 1C, after mutation we can have the tree of figure 1D.

Finally we get the best individual (composite operator) by applying selection operation.

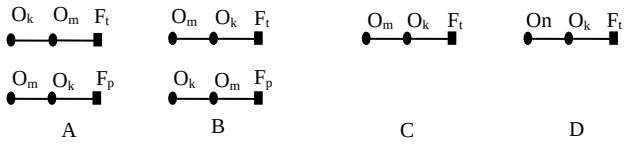


Fig. 1. GP operations are shown graphiclly

The steps of Steady-state genetic programming are folowing:

1) For gen = 1 to N do // N is the number of generation.
2) Keep the best composite operator in P.
3) Repeat.
4) Select 2 composite operators from P based on their fitness values for crossover.
5) Select 2 composite operators with the lowest fitness values in P for replacement.
6) Perform crossover operation and let the 2 offspring replace the 2 composite operators selected for replacement.
7) Execute the 2 offspring and evaluate their fitness values.
8) Until crossover rate is met.
9) Perform mutation on each composite operator with probability of mutation rate and evaluate mutated composite operators.
10) After crossover and mutation, a new population P' is generated
11) Let the best composite operator from population P replace the worst composite operator in P' and let P = P'.
12) if the fitness value of the best composite operator in P is above fitness threshold value then stop

IV. Program System Parts

The software system is developed using steady-state genetic programming. The software system consists of two phases: training (see figure 2) and testing (see figure 3).

Training image is an input image that includes an object of a type. For this type a class should be obtained: a composite operator and a template image. The block "Feature extractor" is a collection of operations for getting the set of terminals. Steady-state genetic programming is used in the block "Genetic programming". "Ground Truth" image is created manually from the input image.

Training part: The composite operator is synthesized for a definite class of images in the training part. Then segmented template image is obtained from the training image by means of the composite operator. The segmented [8, 9] template image and the composite operator are saved as identifiers of class to which the training image belongs. As a result we will have one or more composite operator/operators and one or more template image/images for a definite image. For different classes class identifiers are saved separately.

Testing part: Our system recognizes an image for which the system has a composite operator and a template image in the testing part. Recognition is performed in the following way: the composite operators from all the classes are run on the testing image (separately and parallelly), afterwards the results are segmented as binary images. Each class will have its segmented binary images. Then will the fitness of the resulting images for all classes is calculated separately: in the fitness function as Ground Truth image will be the template image, which corresponds to the composite operator.

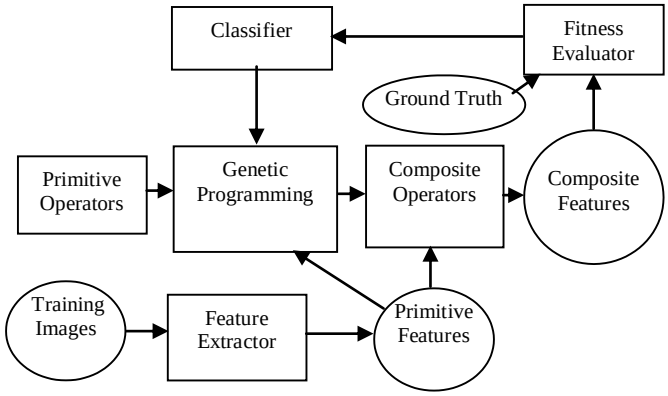


Fig. 2. Learning part diagram of programming system

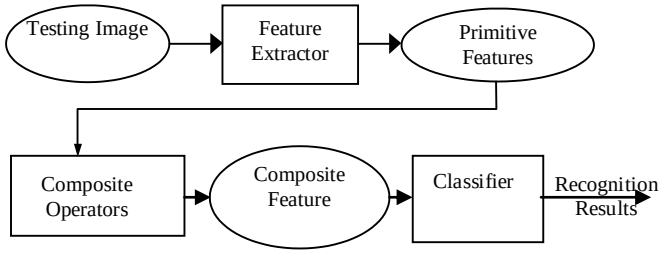


Fig. 3. Testing part diagram of programming system

V. Experiments

Two experiments are performed for recognizing different manmade objects in SAR images.

A. Experiment 1:

Composite operators and segmented binary template images for two kinds of objects (car and helicopter from A SAR image from MSTAR Image database) are found. A car from the SAR image (see figure 4a) is used to synthesize the composite operator, as Ground Truth image figure 4b is used.

Fig. 4. A car image region from real SAR image. a) Real image of car, b) Ground Truth, c) Binary segmented image after composite operator

Synthesized composite operator is flowing:

ADD(PFIM0,FMAX2(FDIV(ADD(FLOG(FDIVC(PFIM0)), FADDC(FDIV(FMUL(PFIM5, PFIM5), ADD(PFIM13, PFIM0)))), ADD(FSUBC(FADDC(PFIM10)), PFIM0)), FLOG(FDIV(PFIM10, FSUBC(FSUBC(FSUBC(PFIM0)))))))

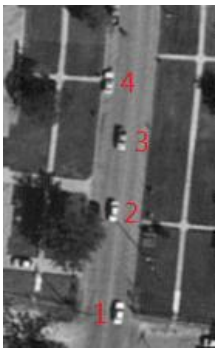


Fig. 5. Cut region from real SAR image

Fig. 6. Cut region from real SAR image

The result of composite operator is figure 4c. Figure 6 A is saved as class template to recognize cars like on figure 4a on the same azimuth angle. The composite operator was applied on car 1 in figure 5 to obtain class template for “car 1” like cars on the same azimuth angle, the result of which was figure 6 B image. Using the composite operator and 6 B class template the recognition rates of car 1, car 2 and car 4 were 0.95, 0.7 and 0.69 respectively. Car 3 was recognized as another object. The composite operator for recognition helicopter is the following:

ADD(PFIM14, ADD(FADDC(FMAX2(FSQRT(PFIM12), FMIN(FMIN(FDIV(PFIM12, PFIM12))))), FSUB(PFIM7, FMEAN(FMAX2(PFIM2, PFIM14))))), which is obtained by using figure 7 helicopter 1 as training image. As Ground Truth is used figure 8 A and the resulting image of the composite operator is shown in figure 8 B, which is saved as template

image for helicopter on the same azimuth angle. The composite operator and 8 B template image are used to recognize helicopters from fig. 7.1, 7.2, and 7.3. The results of recognition were 0.73, 0.7 and 0.73 respectively.

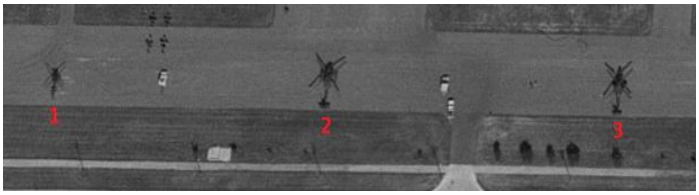


Fig. 7. Cut region from real SAR image

A B

Fig. 8. Cut region from real SAR image

B. *Experiment 2:*

The experiment was done on five types of object images from [2].

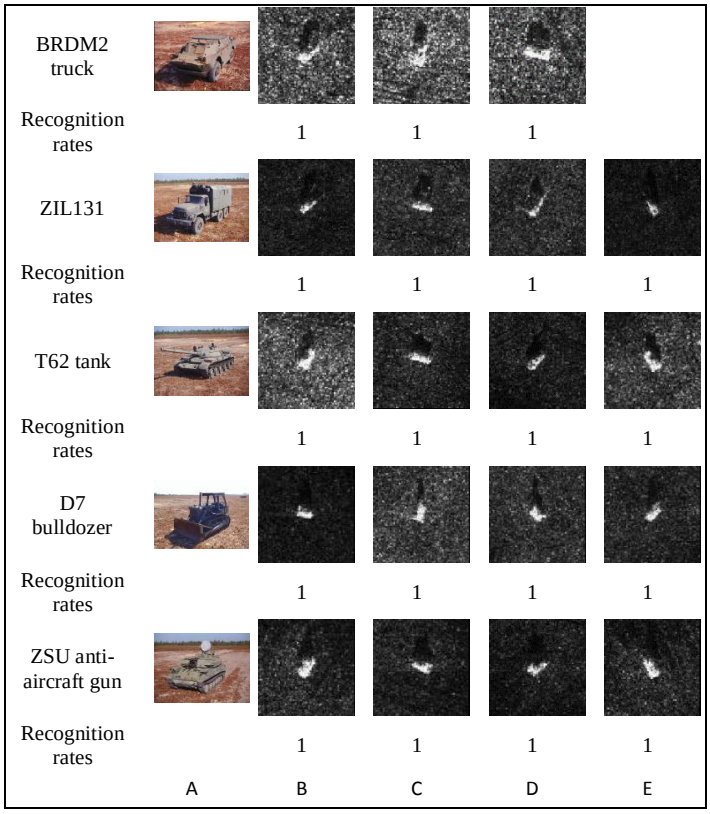


Fig. 9. SAR images of five types of objects and its recognition rates

Figure 9 shows one optical and four SAR images of each object under 15°-depression angle and various azimuth angles between 0° and 359°. One of the four images for each object are used as training image in our program system to get class identifiers for these five types of objects. Composite operators are synthesized for each class (type of object) separately. Class template images are obtained for each SAR images in figure 9 B to E, using those composite operators. As we already mentioned, our system is able to recognize the same object in range $-2^0 <= $ template azimuth angle$<=2^0$ by means of a class template image. The recognition rates of SAR images from figure 8 are equal to 1. Various experiments results are given in figure 10. As test images for our program system are given images obtained from SAR images from figure 8, rotated -2^0 to 2^0 angels. We achieve the same results in two cases (results are given in figure 9):

1) When our system recognizes these five objects.

2) When our system recognizes these five objects and added cars and helicopter objects.

The best experiment results from [2] are given in table 2. Experiments are done for three and five types of objects. We can see that results for five objects are worse than the results of three objects.

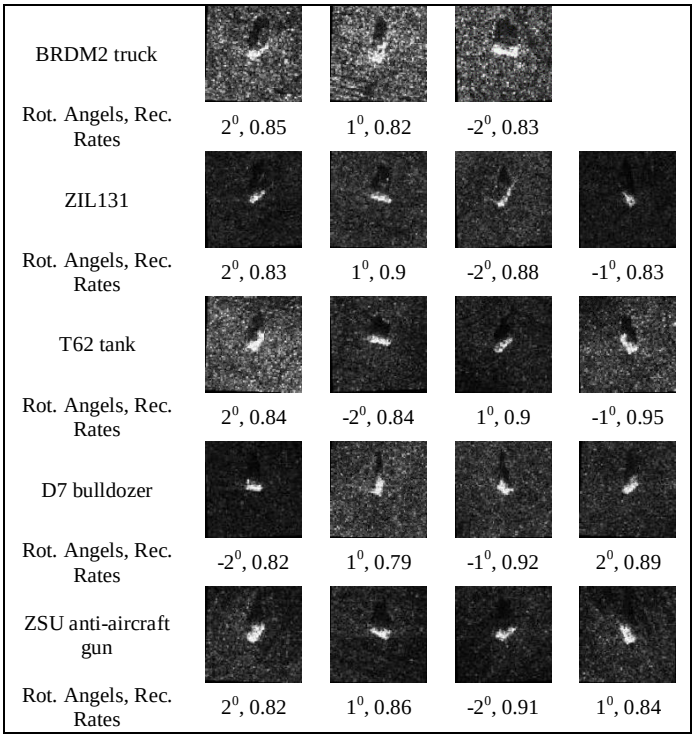


Fig. 10. Rotated SAR images of five types of objects and its recognition rates

TABLE II. EXPERIMENTS RESULTS FROM [2]

Experiment results for tree type objects			Experiment results for five type objects		
Runs	Training	Testing	Runs	Training	Testing
1	1.0	0.967	1	0.929	0.824
2	1.0	0.971	2	0.926	0.816
3	1.0	0.976	3	0.929	0.803
4	1.0	0.964	4	0.941	0.845
5	1.0	0.967	5	0.931	0.809
6	0.995	0.979	6	0.911	0.809
7	1.0	0.979	7	0.903	0.832
8	0.995	0.95	8	0.943	0.842
9	1.0	0.964	9	0.94	0.847

10	0.995	0.983	10	0.909	0.823
Average	0.999	0.97	Average	0.926	0.825

VI. CONCLUSION

A program system, which detects and recognizes objects in SAR images, was presented in this paper. A new approach was used for recognizing objects by using genetic programming to achieve the same good recognition rate, despite the increase of the quantity of object types. One of its benefits was the usage of the primitive set of terminals and the simple genetic programming algorithm.

REFERENCES

[1] http://en.wikipedia.org/wiki/Outline_of_object_recognition

[2] Bir Bhanu and Yingqiang Lin, "Object Detection in Multi-modal Images Using Genetic Programming," University of California, Riverside, CA, 92521, USA, October 2003.

[3] Riccardo Poli, William B. Langdon, Nicholas F. McPhee, John R. Koza, "A Field Guide to Genetic Programming," March 2008

[4] Nerses Safaryan, "Detection and Classification of Objects by Applying Genetic Programming", Mathematical Problems of Computer Science 32, pp. 101-106, 2009.

[5] Nerses Safaryan, "Generation Of Operators To Identificatify Of Objectss In Digital Images," (in russian), XXXVI Gagarin readings Moscow Volume 4, pp. 131-132, April 6-10, 2010.

[6] Nerses Safarayan, Hakob Sarukhanyan, "Learning Features By Means Of Genetic Programming For Object Recognition," Conference Computer Science and Information Technologies, 26 - 30 September, 2011, pp 382-385, Yerevan, Armenia.

[7] D. Howard, S. C. Roberts, and R. Brankin, "Target detection in SAR imagery by genetic programming," Advances in Engineering Software, Vol. 30, No. 5, pp. 303-311, Elsevier, May 1999.

[8] J. Shi and J. Malik. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. and Machine Intell, 2000.

[9] Bhanu and S. Fonder, "Learning-integrated interactive image segmentation," chapter in Advances in Evolutionary Computing–Theory and Application, A. Ghosh and S. Tsutsui (Eds.), pp. 863–895, Springer-Verlag, 2003.

[10] Paul L. Rosin, Javier Hervas, "Image Thresholding For Landslide Detection By Genetic Programming," Proceedings of the First International Workshop on Multitemporal Remote Sensing Images, 13-14 Septembe, 2001 University of Trento, Italy

Route Optimization in Network Mobility

Classification & Comparison

Md. Hasan Tareque
Department of Computer Science and Engineering
IBAIS University
Dhaka, Bangladesh

Ahmed Shoeb Al Hasan
Department of Computer Science and Engineering
Bangladesh University of Business & Technology
Dhaka, Bangladesh

***Abstract*—NEtwork MObility (NEMO) controls mobility of a number of mobile nodes in a comprehensive way using one or more mobile routers. To choose a route optimization scheme, it is very important to have a quantitative comparison of the available route optimization schemes. The focus of this paper is to analyze the degree of Route Optimization (RO), deploy-ability and type of RO supported by each class in general. The comparison shows the differences among the schemes in terms of issues, such as additional header, signaling and memory requirement. We classify the schemes established on the basic method for route optimization, and equal the schemes based on protocol overhead, such as header overhead, amount of signaling, and memory requirements. Lastly the performance of the classes of different schemes has to be estimated under norms such as available bandwidth, topology of the mobile network and mobility type.**

Keywords—Delegation; Hierarchical; Source Routing; BGP-assisted; Network Mobility; Route Optimization

I. INTRODUCTION

The demand for wireless connectivity is rising now a days that it used both static and mobile IP-enabled devices. In the future, it may be common for several devices which are connected in a Local Area Network to move together.

Existing Internet is not designed to handle mobility due to IPs location-based addressing scheme where IP addresses are tied to geographical areas. A host moving between networks in different geographical areas needs to obtain a new IP address, and therefore, communication may become inefficient while maintaining reachability and session continuity. To overcome the inefficiency of current IP addressing, Internet Engineering Task Force (IETF) designed solutions such as Mobile IP (MIP) [1] and MIPv6 [2] to support mobility of a host. A summary of some of the host mobility protocols (including MIP and MIPv6) can be found in [3].

Handling mobility of a number of devices in a moving LAN or PAN using host mobility protocols for each device growths signalling overhead during handoff, power consumption and manageability. IETF developed NEtwork Mobility (NEMO) where one or more routers, called mobile routers, manage the mobility of all the hosts in a network. NEMO supports nested mobile network. IETF protracted MIPv6 to design NEMO Basic Support Protocol (NEMO BSP) [4] to grip network mobility, where hosts in a mobile network are reachable through a home agent.

II. NEMO

NEMO works by moving the mobility functionality from Mobile IP mobile nodes to a mobile router. The router is able to change its attachment point to the Internet in a manner that is transparent to attached nodes. Reduced Signalling, Increased manageability, Reduced power consumption, Conservation of bandwidth, Network Mobility & Nested Mobile network are the main advantages of NEMO.

There are also some drawbacks in NEMO like inefficient routing which will increase end-to-end delay, bandwidth inefficiency and fragmentation. Due to encapsulation increase header overhead. Handoff latency may also rises due to NEMO.

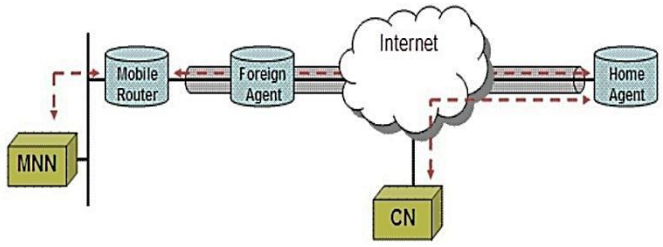


Fig. 1. Basic Idea of NEMO

In fig. 1, the simple idea of NEtwork MObility (NEMO) is been illustrated.

A. Basic Design of NEMO

For better understanding the basic NEMO structure there are some terminologies:

- Mobile Router = MR.
- Mobile Network Node = MNN.
- Top Level Mobile Router = TLMR.
- Access Router = AR.
- Binding Update = BU.
- Binding Acknowledgement = BA.

There are several types of MNNs:

- Local Fixed Node (LFN): Nodes which do not move with respect to the mobile network.
- Local Mobile Node (LMN): Nodes which usually reside in the mobile network but can move to other networks.

- Visiting Mobile Node (VMN): Nodes which belong to another network but is currently attached to the mobile network.
- MR: An MNN can act as an MR to form a nested mobile network.
- Mobility Capable Nodes (MCNs): LMNs, VMNs and MRs implement mobility protocols; these are referred as MCNs.
- Home Network: The network to which a mobile network is usually connected.
- Home Agent (HA): An MR is registered with a router, called Home Agent, in its home network.
- Correspondent Node (CN): A node that communicates with MNNs.
- Home Address (HoA): Address through which TLMR is reachable in its home network. In fig. 2, the mobile network under MR1 is nested under TLMR's mobile network; MR1's mobile network thus has a nesting level of one.

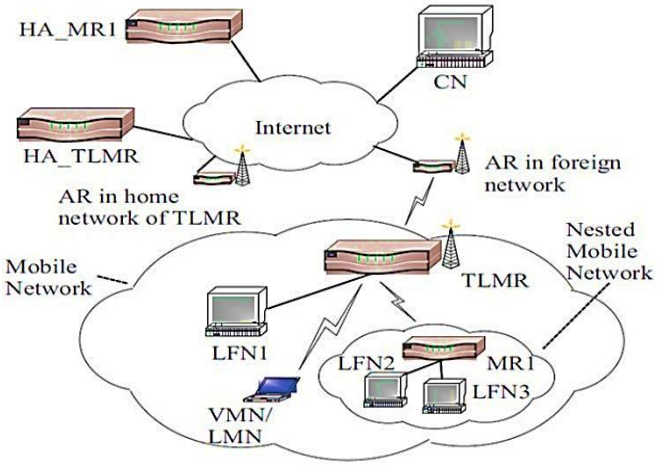


Fig. 2. Architecture of NEMO showing one level of nesting

- Care-of-Address (CoA): When the TLMR moves to a foreign network (any network other than home network), it obtains a new address from the foreign network.

An excellent style manual for science writers is [7].

III. Routing Optimization (RO)

Route Optimization (RO) is solving the problem of inefficient route and header overhead. The basic principle of RO is to enable packets to directly reach the mobile network by avoiding multiple tunnels through home agents.

To trade off the gain of RO with the performance and applicability, several schemes have been proposed. In this section, we present the RO schemes and their challenges in and issues.

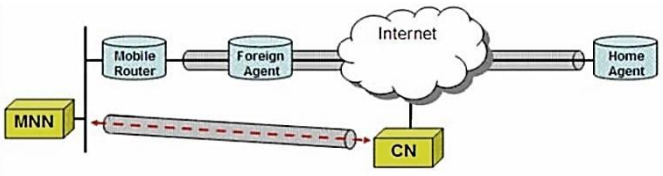


Fig. 3. Routing Optimization (RO)

In fig. 3, Shows how route optimization is been done in NEMO.

A. Challenges in RO

RO requires bypassing the HAs when packets are sent between CN and MNNs. Bypassing has given rise to the following two major challenges which have to be addressed by RO schemes:

- How can a packet destined to an MNN reach the TLMR attached to the foreign network to which the MNN is attached (directly or indirectly)?
- How is a packet routed inside the mobile network after reaching TLMR?

The challenge of RO in intra mobile network case (Intra RO) is how to route packets between two MNNs without letting the packet outside the mobile network. Intra RO [5] is included because they also optimize route for communication between a CN and an MNN.

B. Issues in RO

There are several issues [6] that were raised in addition to header overhead and Intra RO those issues are given below.

- Signaling

When a mobile network moves, only the MR to which the movement is visible needs to perform signaling with its HA. Signaling packets competes with data packets for bandwidth not only inside the mobile network but also in the Internet.

- Memory requirement

The schemes have to maintain various state information, regarding the route and CN-MNN pairs. Example: small sensors and PDAs.

- Degree of RO

In an effort to trade off issues, such as signaling, some schemes allow one or two levels of tunneling or some non-optimality in the route between a CN and an MNN.

- Header overhead

Header overhead is the additional information that is put into the header for RO. Header overhead consumes bandwidth and increases chance of fragmentation.

- Intra RO

Route optimization between two MNNs within a mobile network is called Intra RO. With a focus on optimizing route between a CN and an MNN, some of the schemes do not consider Intra RO.

- Deploy-ability

Changes in mobility entities are tolerable because they are going to be introduced in the existing infrastructure if NEMO support is required. Changes in functionalities in hosts and routers in the existing infrastructure may not be easily applicable resulting in concern about deploy-ability issue.

- Location management

Location management is tracking the location of an MNN to ensure reachability and session continuity. In NEMO BSP and some RO schemes, location management is performed by HA.

- Location transparency

In NEMO BSP, MNNs except MRs and CNs are transparent.

IV. RO SCHEMES

There are several RO schemes have been proposed to resolve the issues in RO. Based on approach used, the various RO schemes that have been proposed can be generally classified as:

- Delegation
- Hierarchical
- Source Routing
- Border Gateway Protocol (BGP)-assisted

The basic principle of each class, and a description and comparison of the schemes are discussed throughout the next section.

A. Delegation Schemes

In this class, prefix of the foreign network is delegated inside the mobile network. MCNs contain CoAs from the prefix and send BUs to respective HAs and CNs. So, any packet from CN, addressed to CoA, reached the foreign network without going through HAs.

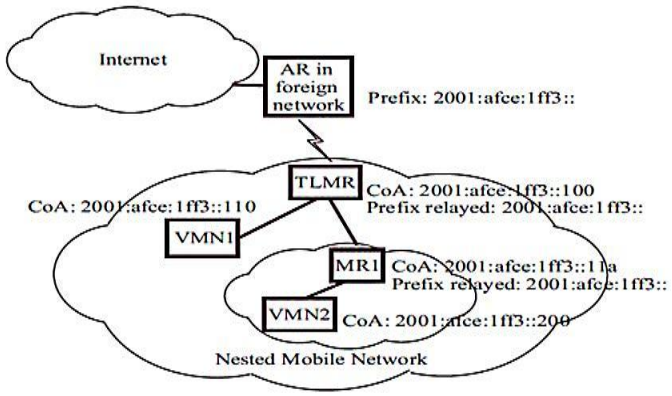


Fig. 4. Delegation approach for route optimization

In Fig. 4, prefix 2001:afce:1ff3:: is relayed by TLMR inside its mobile network. VMN1 and MR1 obtains CoA 2001:afce:1ff3:110 and 2001:afce:1ff3:11a, respectively; and MR1, in turn, relays the prefix inside its network.

The concept of prefix delegation is simple [7], and provides optimal route with low header overhead at the cost of sacrificing location transparency. By, sending BU to CN requires additional signaling along with requirement of protocol sup- port from CN, making the schemes difficult to execute. The schemes do not concentrate on Intra RO.

B. Hierarchical Schemes

In the hierarchical class, a packet, rather than traveling through all HAs, reaches the foreign network either from MNNs HA (first HA) or traveling only through HA of MNN and TLMR. Unlike delegation-based approach, an MR does not send its CoA to CNs. Rather; an MR sends TLMRs CoA or HoA to HA. CNs use MNNs HoA to send packets to an MNN. Packets, sent by CN to MNN, reach MNNs HA that tunnels the packets to TLMRs CoA or HoA. Packets, tunneled to CoA, directly reach the foreign network, whereas packets, tunneled to HoA, reach TLMRs HA that tunnels packets to TLMR. On reaching TLMR, packets are routed to MNN by MRs that maintains a routing table containing the mapping of MNNs prefix to next hop MR.

In Fig. 5, the abstract view of the hierarchical class is shown. TLMR CoA is passed to HA MR1 and HA VMN by MR1 and VMN, respectively. Also, MR1 and VMN send their CoAs to TLMR to enable forwarding inside the mobile network. Therefore, a packet sent to VMN will first reach HA VMN that tunnels the packet to the TLMR for forwarding towards the VMN. Thus, communication route is divided into two parts.

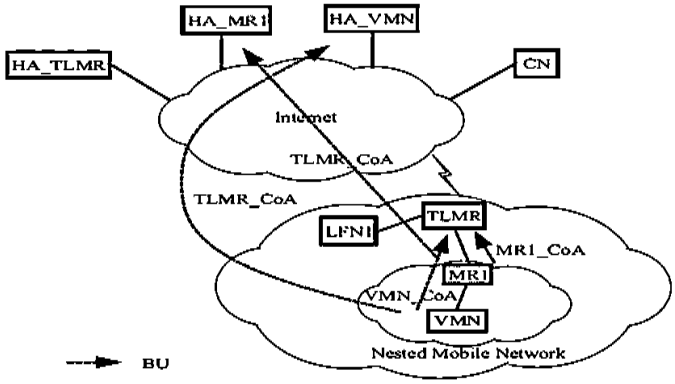


Fig. 5. Hierarchical approach for route optimization

- The route between TLMR and HA VMN.
- The route from the TLMR to VMN.

asAt least one tunnel always exists between the TLMR and HA VMN. The route between CN and MR1 is similar to that between CN and VMN.

The schemes in this class mainly differ in the use of TLMRs CoA or HoA for tunneling, techniques to convey TLMRs address to MRs, and routing of packets inside mobile network resulting in differences in signaling, memory requirement and degree of RO. Moreover, depending on the use of HoA or CoA of the TLMR, the number of tunnels used for communication differs among the schemes; number of

tunnels affects degree of RO and header overhead. In addition, location management entities also vary among the schemes.

The schemes have the disadvantage of packets going through one or two tunnels, resulting in near optimal route and header overhead.

C. Source Routing Schemes

In this class, RO is achieved by sending the CoAs of MRs to the CN which, like source routing, inserts the CoAs in the packet header to reflect the nesting structure of the MRs. This however, results in increased header overhead. Packets from the CN reach TLMR in an optimal route (without going through HAs); routing within the mobile network is done using the CoAs in the packet header. Memory requirement for routing entries is low because each MR needs to keep track of only the attached MRs as next hop. Schemes in this class notify CN about the CoAs of MRs in various ways that will be detailed in the descriptions of the schemes. Notification of CoAs to CNs sacrifices location transparency and increases signaling. Methods of notifying the CN result in differences in signaling and overheads. Moreover, the schemes also have different memory requirement for routing packets inside the mobile network.

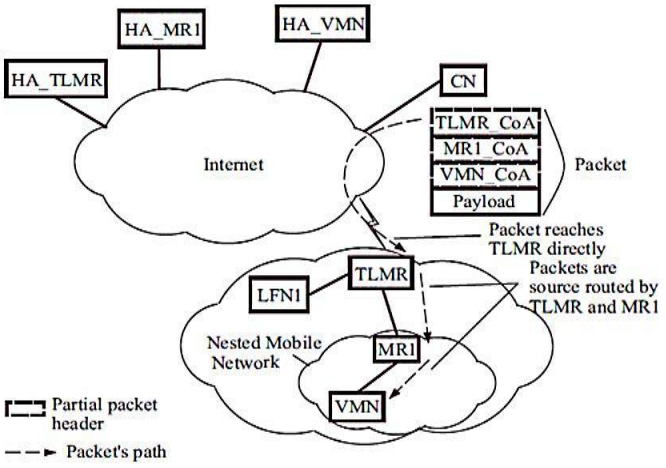


Fig. 6. Source routing approach

Fig 6. Shows the basic principle of the source routing approach where CoAs of TLMR, MR1 and VMN are inserted in packets. Packets, on reaching TLMR, are source routed (using the CoAs) inside the mobile network by TLMR and MR1.

D. Border Gateway Protocol (BGP)-assisted Schemes

The schemes in this class rely on BGP [8] for mobility management. When the mobile network moves, BGP routers are updated to make necessary changes in the routing tables by making forwarding entries for the prefix of the mobile network. Information regarding the change of route of the mobile network is signaled to few routers that exchange the information with peers using existing routing protocols in the Internet. Therefore, routers contain routing entries to route packets to the mobile network irrespective of its location, and are responsible for location management. Schemes in this class mainly differ in the number of external BGP updates generated, and incurring other overheads for managing Intra RO.

An abstract view of the approach used in this class has been shown in Fig. 7. When the TLMR joins the AR in the foreign network, AR injects a BGP update that maps TLMRs prefix (1:3:1::) to ARs address (1::2). BGP router3 in ARs network updates its peers (BGP router1 and BGP router2), accordingly. Therefore, packets sent by CN will reach a BGP router in its network, and will be forwarded to the appropriate BGP routers network where the mobile network resides.

The major advantage of the schemes in this class is the use of no new entity for mobility management. Moreover, CNs are transparent to the change location (managed by BGP routers) of the MNNs. On the other hand, these schemes will produce a storm of updates.

This trade of also requires packets always traveling through one or more of some designated routers resulting in near optimal route.

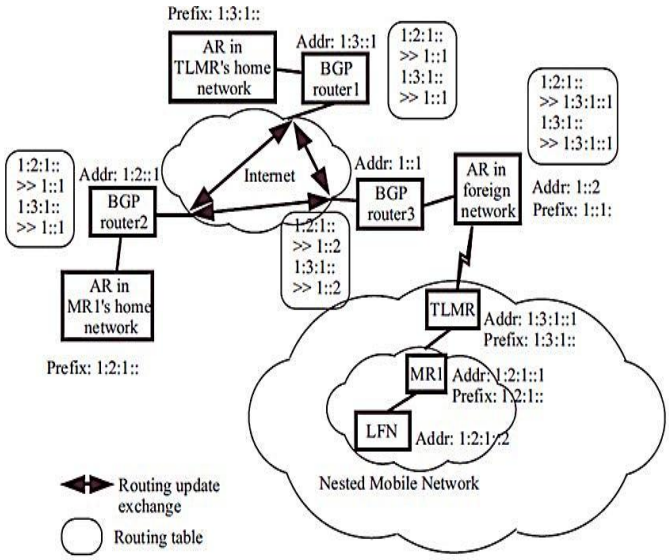


Fig. 7. BGP-assisted approach.

E. Miscellaneous Schemes

This section includes RO schemes that do not fall into any of the previous classes described. The techniques, used for RO in the schemes presented in this section, are different than the basic techniques used for RO.

a) Optimized Route Cache (ORC) – based: An approach [9] where the MR sends BU to a router in the CNs network, and to the MR attached above (parent MR). A major disadvantage of ORC is that it optimizes route for only one level of nesting. Although route from CNs to MNNs is similar to that in hierarchical class for one level of nesting, it is different when the nesting level increases. In hierarchical class, TLMRs HoA or CoA is not conveyed to the nested MRs. Therefore, we have placed this scheme separately in this section

b) Recursive BU (RBU) – based: A RO scheme [10] where BUs, sent by MRs to CN, are used to recursively process the

binding table at CN to maintain a route to TLMR. Memory requirement for routing will be low when routes are discovered dynamically. Signaling in this scheme is high.

c) AODV − based: The route between a HA and an MR is established using AODV protocol [11]. The scheme appears to be very simple; yet, it requires all routers in the Internet, and HA to support AODV resulting the scheme difficult to deploy. Moreover, the scheme involves one tunnel for communication along with overhead of burst of messages in the Internet during handoff due to broadcast of AODV messages. Although AODV is a protocol for Ad hoc networks, we do not include AODV-based scheme in delegation class under Ad hoc-based scheme due to the following reason: The basic principle used in Ad hoc- based scheme is to obtain a CoA from the foreign network prefix contrasting the obtaining of CoA from MRs prefix in the AODV-based scheme.

TABLE I. A COMPARISON AMONG DIFFERENT CLASSES

Class	Degree of RO	Intra RO	Signalling	Header Overhead	Deployability	Location Transparency
Delegation	Optimal	No	High	Low	Difficult	No
Hierarchical	Near optimal	Yes	Low	Medium	Easy	Yes
Source routing	Optimal	No	High	High	Difficult	No
BGP-assisted	Near optimal	Yes	Low	Low	Difficult	Yes

V. ANALYSIS

In this section, we perform comparison of different schemes which was discussed above.

- The comparisons show that hierarchical scheme are easier to deploy, and also supports efficient intra mobile network communication in the wired network.
- Delegation-based and BGP-assisted schemes suites the client-server type communication that prevails in the existing Internet.
- Delegation approach is simple, do not introduce any additional overhead on Internet routing, and optimize route completely.
- BGP-assisted approach supports Intra RO, and requires fewer supports from infrastructure.
- Source routing approach is not suitable for mobile networks having higher nesting levels due to higher header overhead that consumes bandwidth which in wireless environment.
- There is a comparison shown in Table I among different schemes [Delegation, Hierarchical, Source Routing, BGP- assisted]

VI. FUTURE RESEARCH

Although a considerable amount of research has been carried out in NEMO, there are still lots of issues where future research can be possible.

- Further Research is required to determine the performance of the RO schemes with change in network topology.
- Most of the RO schemes incur additional signaling over bandwidth limited wireless channels. This contradicts one of the initial objectives of NEMO as a scheme to reduce signaling over wireless channels by letting the mobile router carry out the signaling on behalf of all the nodes. To reduce the signaling, a constant level (one or two) of tunneling can be allowed.

Again update for all the CNs and HAs can possible with a single BU. This is done by letting the CNs and HAs join a multicast group when they join the mobile network. RO schemes can be analyzed to find suitable schemes based on the architecture of the mobile network, availability of bandwidth, and mobility pattern. To reduce the signaling, a constant level of tunneling can be allowed. This has been done in some of the schemes in hierarchical class, and can be adopted in other schemes dynamically on ad hoc basis.

- Considering handoff performance [12] [13] along with any RO scheme is important.
- Security threats [14] also need to be considered in conjunction with RO.

VII. CONCLUSION

The mobile network that supports NEMO is foreseeable in mobile platforms such as car, bus, train, air-plane, etc. The use of NEMO BSP [4] gives more assistances than MIPv6 [2] in mobile platforms. The restrictions of NEMO can be addressed by RO schemes. Although RO schemes address the problems of NEMO, ensuring QoS is great challenge in various internet applications.

In this paper, we present classification and comparison among the RO schemes for NEMO. The number of RO schemes reported in this article indicates the exhausting and diverse efforts for RO, and therefore, requires a quantitative evaluation of the RO schemes to determine their suitability and adaptability to the existing Internet infrastructure.

The comparison among the schemes within each class reveals the differences among the schemes in more depth. Moreover, signaling and memory requirement depend on the number and types of MNNs in the mobile network, and

therefore, might guide the selection of the schemes. Hence, the evaluation under various parameters is required to determine the suitability of the schemes. To apply the RO schemes to real-world applications and enable their wide deployment, protocol overheads, such as header overhead and signaling need to be reduced.

The internet applications can be classified into real-time and non-real-time applications. The real-time applications are real-time interactive audio and video applications; one-to-many streaming of real-time audio and video applications; streaming of stored audio and video applications. The non-real time applications are file transfer, web access, e-mail, etc. As the QoS (Quality of Service) requirements for each application can vary from each other, suitable selection of existing RO schemes was done. It is obvious that most of the RO schemes are not considered all QoS parameters such as delay, jitter, and bandwidth and packet loss. In future, it is necessary to focus on these QoS parameters. The limitations of each and every scheme can be further studied with respect to QoS requirements.

ACKNOWLEDGMENT

A special thanks to Dr. Md. Shohrab Hossain for giving us a chance to work with this topic. Also a superior thanks goes to Abu Zafar M. Shahriar, Mohammed Atiquzzaman, there work was really helpful.

REFERENCES

[1] C. Perkins, "IP mobility support for IPv4," RFC 3220, Jan 2002.

[2] D. B. Johnson, C. E. Parkins, and J. Arkko, "Mobility support in IPv6," RFC 3775, Jun 2004.

[3] A. Conta and S. Deering, "Generic packet tunneling in IPv6 specifications," RFC 2473, Dec 1998.

[4] V. Devarapalli, R. Wakikawa, A. Petrescu, and P. Thubert, "NEtwork MObility (NEMO) basic support protocol," RFC 3963, Jan 2005.

[5] C. Ng, P. Thubert, M. Watari, and F. Zhao, "Network mobility route optimization problem statement," RFC 4888, Jul 2007.

[6] C. Ng, F. Zhao, M. Watari, and P. Thubert, "Network mobility route optimization problem statement," RFC 4889, Jul 2007.

[7] K. Lee, J. Park, and H. Kim, "Route optimization for mobile nodes in mobile network based on prefix delegation," IEEE 58th Vehicular Technology Conference, pp. 2035–2038, Oct 6-9 2003.

[8] Y. Rekhter, T. Li, and S. Hares, "A border gateway protocol 4 (bgp-4)," RFC 4271, Jan 2006.

[9] R. Wakikawa, S. Koshiba, K. Uehara, and J. Murai, "Orc: Optimize route cache management protocol for network mobility," 10th International Conference on Telecommunications, pp. 1194–1200, Feb 23 - Mar 1 2003.

[10] H. Cho, E. K. Paik, and Y. Choi, "Rbu+: Recursive binding update for end-to-end route optimization in nested mobile networks," 7th IEEE International Conference on High Speed Networks and Multimedia Communications, pp. 468–478, Jun 30 - Jul 2 2004.

[11] R. Cuevas, A. Cabellos-Aparicio, A. Cuevas, J. Domingo-Pascual, and A. Azcorra, "fp2p-hn: A p2p-based route optimization architecture for mobile ip-based community networks," Computer Networks, pp. 528–540, Mar 2009.

[12] H. Petander, E. Perera, K. Lan, and A. Seneviratne, "Measuring and improving the performance of network mobility management in IPv6 networks," IEEE J. Sel. Areas Commiun, pp. 1671–1681, Sep 2006.

[13] H. K. Ryu, D. H. Kim, Y. Z. Cho, K. W. Lee, and H. D. Park, "Improved handoff scheme for supporting network mobility in nested mobile networks," International Conference on Computational Science and Its Applications, pp. 378–387, May 9-12 2005.

[14] S. Jung, F. Zhao, and H. Kim, "Threat analysis on network mobility (NEMO)." Sixth International Conference on Information and Communications Security, Oct 27-29 2004.

Survey: Risk Assessment for Cloud Computing

Drissi S.
Computer Lab and Renewable Energy Systems (CLRES)
University Hassan II Aïn Chock. ENSEM
Casablanca, Morocco

Houmani H. and Medromi H.
Computer Lab and Renewable Energy Systems (CLRES)
University Hassan II Aïn Chock. ENSEM
Casablanca, Morocco

***Abstract*—with the increase in the growth of cloud computing and the changes in technology that have resulted a new ways for cloud providers to deliver their services to cloud consumers, the cloud consumers should be aware of the risks and vulnerabilities present in the current cloud computing environment. An information security risk assessment is designed specifically for that task. *However, there is lack of structured risk assessment approach to do it.* This paper aims to survey existing knowledge regarding risk assessment for cloud computing and analyze existing use cases from cloud computing to identify the level of risk assessment realization in state of art systems and emerging challenges for future research.**

Keywords—cloud computing; risk; risk assessment approach; survey; cloud consumers

I. INTRODUCTION

With the advancement in cloud technologies and increasing number of cloud users, businesses also need to keep up with the existing technology to provide real business solutions [1]. In addition, predictions for growth indicate massive developments and implementations of cloud computing services, including that the cloud computing services market is likely to reach between $150 billion in 2014 [29-30] and $222.5 billion in 2015 [31]. From the business perspective, cloud computing becomes one of the key technologies that provide real promise to business with real advantages in term of cost and computational power [2]. In spite of the advancement in cloud technologies and increasing number of cloud users, Cloud computing being a novel technology introduces new security risks [22] that need to be assessed and mitigated. consequently, assessment of security risks [17] is essential , the traditional technical method of risk assessment which centers on the assets should give way to the business focused on the specific nature of cloud computing and on the changes in technology that have resulted a new ways for cloud providers to deliver their services to cloud consumers.

The major contributions of this survey can be summarized as follows:

a) We investigate the existing knowledge regarding risk assessment for cloud computing.

b) Further, we also present a risk assessment requirement that can be used by a prospective cloud consumers to assess the risk in cloud computing.

The rest of the paper is organized as follows: Cloud computing and concepts of risk assessment are summarized in Section 2. In Section 3, we are investigated the major paradigms of risk assessment in cloud computing. New researches requirements for risk assessment in cloud computing environment are discussed in Section 4. Finally, the survey concludes with the open challenges of risk assessment in cloud computing environment in Section 5.

II. FUNDAMENTAL CONCEPTS

A. Cloud computing

In literature, there are many definitions for cloud computing. The National Institute of Standards and Technology(NIST) [4] defines cloud computing as ‘‘a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction’’. European Community for Software and Software Services (ECSS) [5] explains it as the delivery of computational resources from a location other than your current one.

Cloud can be categorized into three delivery models classified according to their uses; Cloud Software as a Service (SaaS), Cloud Platform as a Service (PaaS) and Cloud Infrastructure as a Service (IaaS). Cloud Software as a Service (SaaS) which deliver software over the Internet (e.g. Salesforce CRM, Google Docs, etc), Cloud Platform as a Service which mainly offer virtualized execution environments to host Cloud services (e.g. Microsoft Azure, Force and Google App engine) and Cloud Infrastructure as a Service which provide virtualized computing resources as a service (e.g. Amazon EC2 and S3, Terremark Enterprise Cloud, Windows Live Skydive and Rackspace Cloud).

Four deployment models have been identified for cloud architecture solutions: Private cloud: a cloud platform is operated for specific organization, Community cloud: The cloud infrastructure is shared by several organizations and supports a specific community that has communal concerns, Public cloud: a cloud platform available to public users to register and uses the available infrastructure. Hybrid cloud: a private cloud that can composite two or more clouds (private, community or public).

B. Risk assessment

“Risk in itself is not bad, risk is essential to progress, and failure is often a key part of learning. But we must learn to balance the possible negative consequences of risk against the potential benefits of its associated opportunity” [28].

Risk management refers to a coordinated set of activities and methods that is used to direct an organization and to control the many risks that can affect its ability to achieve

objectives. According to the introduction to ISO 31000 2009, the term risk management also refers to the architecture that is used to manage risk [6]. Risk assessment is one step in the process of risk management.

Risk assessment is the process of identifying the security risks to a system and determining their probability of occurrence, their impact, and the safeguards that would mitigate that impact. The main objective of risk assessment is to define appropriate controls for reducing or eliminating those risks.

Generally there are four steps of risk assessment. The four steps are as follow [7]:

1) Threat Identification

This first step identifies all potential threats to the system. It allows identifying the potential threat sources and develops a list of a threat statement that is potential threat sources that are applicable to the system.

2) Vulnerability Identification

In the second step, the goal of vulnerability identification is to develop a list of system vulnerabilities (flaws or weaknesses) that could be exploited by the potential threat-sources.

3) Risk Determination

In the third step, the purpose of risk determination is to assess the level of risk to the system.

4) Control Recommendation

In the fourth step, the goal is to purpose some controls that could mitigate or eliminate the identified risks, as appropriate to the system organization's operations, are provided. The goal of the recommended controls is to reduce the level of risk to the system.

Risk analysis methods are generally divided into qualitative analysis and quantitative analysis:

Quantitative Risk Methodologies**:** Although there are many well-developed industries that use quantitative risk, it is not commonly used in information technology. In fact, it is very rare indeed. However, risk methodologies can be partially quantitative and partially qualitative. It is the position of this author however to categorize all of the major methodologies as essentially qualitative because none of them can produce ALEs that can credibly be used to measure specific costs versus benefits as quantitative risk analysis should. They instead provide a more general sense of cost versus benefit despite sometimes having aspects which are predominantly quantitative, such as incident statistics [20].

Qualitative Risk Assessments**:** approach describes likelihood of consequences in detail. This approach is used in events where it is difficult to express numerical measure of risk. It is, for example, the occurrence without adequate information and numerical data. Such analysis can be used as an initial assessment to recognize risk [34]. The following are some of the major risk assessment methodologies available today:

· EBIOS [8]
· OCTAVE [9]
· MEHARI [10]

Some are publicly available (e.g. OCTAVE), while others are restricted to members of organizations that are collaborating to create and updated them (e.g. SPRINT). The following are brief descriptions of each of these methodologies.

The method to assess risks is generally composed of the four following steps: thread identification, vulnerability identification, risk determination and control recommendation. These four steps of risk assessment are based on practical experiences in security assessment. These steps come from best practices that have been applied by many organizations for security assessment (e.g. EBIOS, MEHARI and OCTAVE).

The EBIOS [8] (Expression of Needs and Identification of Security Objectives) is a method for assessing and treating risks, which aims to determine the security actions to implement and also expressions safety.

OCTAVE [9] (Operationally Critical Threat, Asset, and Vulnerability Evaluation) is a method of assessment of vulnerabilities and threats on the basis of the operating assets of the company.

MEHARI [10] is a risk assessment method in the context of the security of information systems; this method is designed to meet the needs of each company.

These tools have not been designed specifically for cloud environments. In traditional IT environments, everyone in the business has to go to the IT department to obtain IT related services. However, for cloud computing, the risk assessment become more complex, there are several issues that are likely emerged. Among them is the question of multi-tenancy that means the data may be located at several geographically distributed nodes in the cloud and the control over where the processes actually run and where the data reside.

Existing risk assessment methods and standards (ISO/IEC 27001, ISO/IEC 27005, and EBIOS) are generally focused on structuring the different steps and activities to be performed. Their added value also depends on the knowledge bases of risks [24], [25], [26] and security requirements [24], [26] they require. They are the input to the activities performed. The methodological aspects are thus generally rigorous because, they build on a well defined process and structure to be followed.

III. Literature Review

A. Risk assessment for conventional system

Risk assessment has been discussed by many researches in different area. In [38], a risk assessment method has been discussed for Smartphone; this method describes a method for risk assessment that is tailored for Smartphone. The method does not treat this kind of device as a single entity. Instead, it identifies Smartphone assets and provides a detailed list of specific applicable threats. For threats that use application permissions as the attack vector, risk triplets are facilitated.

The triplets associate assets to threats and permission combinations. Then, risk is assessed as a combination of asset impact and threat likelihood. The method utilizes user input, with respect to impact valuation, coupled with statistics for threat likelihood calculation.

In [36], this paper proposes a method for a probabilistic model driven risk assessment on security requirements. The security requirements and their causal relationships are represented using MEBN (Multi-Entities Bayesian Networks) logic that constructs an explicit formal risk assessment model that supports evidence-driven arguments.

Several quantitative risk assessment methods exist. In [35], they propose a SAEM method which is a cost-benefit analysis process for analyzing security design decisions based on the comparison of a "threat index". However, it is based on some impractical assumptions. In [23] they propose security ontology for organizing knowledge on threats, safeguards, and assets. This work constructs classification for each of these groups and creates a method for quantitative risk analysis, using its own framework. The work does not use known standards or guidelines as an input for its evaluation model, so desired mechanisms and countermeasures have to be defined in the process of risk analysis. Quantitative risk-based requirements are reasoning in [21] uses PACT as a "filter" arranged in series to find out a proportion of likelihood or the impact of risk factor. However, it lacks the ability to represent the impacts among multiple risk factors. The SSRAM model in [3] provides a prioritization that helps in determining how the risks identified will be addressed in different phases of software development. However, it lacks a baseline for systematically identifying potential risks and reasoning about their relationships and interactions in a real operational environment.

In [37], a novel approach is proposed, in which Analytic Hierarchy Process (AHP) and Particles Swarm Optimization (PSO) can be combined with some changes, is presented. The method consists of; firstly, the analytic hierarchy structure of the risk assessment is constructed and the method of PSO comprehensive judgment is improved according to the actual condition of the information security. Secondly, the risk degree put forward is PSO estimation of the risk probability, the risk impact severity and risk uncontrollability. Finally, it gives examples to prove that this method Multi Objectives Programming Methodology (MOPM) can be well applied to security risk assessment and provides reasonable data for constituting the risk control strategy of the information systems security.

B. Risk assessment for cloud computing

In recent years, the principles and practices of risk assessment/management were introduced into the world of utility computing such as Grid and Clouds either as a general methodology [40][41][42][16][43][46] or a focus on a specific type of risk, such as security [45] and SLA fulfillment [44][13].

European Network and Information Security Agency (ENISA) released cloud computing Risk Assessment report, in which ENISA pointed out the advantages and security risks in cloud computing, provided some feasible recommendations and designed a set of assurance criteria to assess the risk of adoptions cloud services [11] [12]. In [13], a quantitative risk and impact assessment framework based on NIST- FIPS-199 [33] (QUIRC) is presented to assess the security risks associated six key categories of security objectives (SO) (i.e., confidentiality, integrity, availability, multi- party trust, mutual audit ability and usability) in a Cloud computing platform. The quantitative definition of risk is proposed as a product of the probability of a security compromise, i.e., an occurring threat event, and its potential impact or consequence. The overall platform security risk for the given application under a given SO category would be the average over the cumulative, weighted sum of n threats which map to that SO category. In addition, a weight that represents the relative importance of a given SO to a particular organization and/or business vertical is also necessary and their sum always adds up to 1. This framework adopts a wide band Delphi method [14], using rankings based on expert opinion about the likelihood and consequence of threats, as a scientific means to collect the information necessary for assessing security risks. The advantage of this quantitative approach of risk assessment is that it enables cloud providers, cloud consumers and regulation agencies the ability to comparatively assess the relative robustness of different Cloud vendor offerings and approaches in a defensible manner. However, the challenge and difficulty of applying this approach is the meticulous collection of historical data for threat events probability calculation, which requires data input from those to be assessed Cloud computing platforms and their vendors. Similar efforts were carried out in [48].

In [15], a risk analysis approach from the perspective of a cloud user is presented to analyze the data security risks before putting his confidential data into a cloud computing environment. The main objectives of this work are to help service providers to ensure their customers about the data security and the approach can also be used by cloud service users to perform risk analysis before putting their critical data in a security sensitive cloud. This approach is based on trust matrix. There is a lack of structured analysis approaches that can be used for risk analysis in cloud computing environments. The approach suggested in this paper is a first step towards analyzing data security risks. This approach is easily adaptable for automation of risk analysis. In [16], a Semi-quantitative BLO-driven Cloud Risk Assessment (SEBCRA) approach that is aware of the Business-Level Objectives (BLOs) of a given Cloud organization is presented. The approach is designed for a Cloud Service Provider (CSP) to improve the achievement of a BLO, i.e., profit maximization, by managing, assessing, and treating Cloud risks. The core concept on which this approach is based is that "Risk Level Estimation for each BLO is proportional to the probability of a given risk and its impact on the BLO in question". Once risk has been assessed, the Risk Treatment sub-process defines potential risk-aware actions, controls, and policies to conduct an appropriate risk mitigation strategies, such as, avoid the risk, by eliminating its cause(s), reduce the risk by taking steps to cut down its probability, its impact, or both, accept the risk and its related consequences or transfer or delegate the risk to external organizations. In an exemplary

experimentation, the risk assessment approach demonstrates that it enables a CSP to maximize its profit by transferring risks of provisioning its private Cloud to third-party providers of Cloud infrastructures. This risk assessment approach can be extended to tackle scenarios where multiple BLOs are defined by a CSP and also work as an autonomic risk-aware scheduler, which will be based on business-driven policies and heuristics that help the CSP to improve its reliability.

In [17], a cloud-based risk assessment as a service is proposed as a promising alternative. Cloud computing introduces several characteristics that challenge the effectiveness of current assessment approaches. In particular, the on-demand, automated, multi-tenant nature of cloud computing is at odds with the static, human process-oriented nature of the systems for which typical assessments were designed. However, the autonomic risk assessment is far away from the light, because the risk assessment is hard task to do. In [18], a framework called SecAgreement (SecAg) is presented, that extends the current SLA negotiation standard, WS-Agreement, to allow security metrics to be expressed on service description terms and service level objectives. The framework enables cloud service providers to include security in their SLA offerings, increasing the likelihood that their services will be used. We define and exemplify a cloud service matchmaking algorithm to assess and rank SecAg enhanced WS-Agreements by their risk, allowing organizations to quantify risk, identify any policy compliance gaps that might exist, and as a result select the cloud services that best meet their security needs.

In [27], they present a methodology for performing security risk assessment for cloud computing architectures in deferent stages (deployment and operation) basing on rules of Bayesian dependencies. The main objective of this paper is to prove how to calculate the relative risk (RR) after cloud adoption (RR=1 do nothing, RR<1 accept risk, RR>1 apply mitigation).

In [32], this paper sums up 8 kinds of threats to security principles, and lists the corresponding factors. Combing with collaborative and virtualization of cloud computing technology and so on, adopting the theory of AHP and introducing the correlation coefficient to analyze the multiple objective decisions, the paper proposes a new information security risk assessment model based on AHP in cloud computing environment. Thus, the objective of this paper is to get the security risk assessment strategies of the information system in the cloud computing environment.

IV. Synthesis and Discussion

Most of the current work is for helping cloud consumers assessing their risk before putting their critical data in a security sensitive cloud. All of these researches have laid a solid foundation for cloud computing. However, they barely established a complete risk assessment approach in consideration of the specific and complex characteristics of cloud computing environment. There were neither a complete qualitative or quantitative risk assessment method for cloud computing. Therefore, there is a need of new risk assessment approach for cloud consumers to check the effectiveness of the current security controls that protect an organization's assets.

TABLE I. RISK ASSESSMENT LIMITATIONS FOR CLOUD COMPUTING

Research paper	Characteristics	
	Stakeholders	***Limitations***
[13]	Cloud providers and cloud consumers	The challenge and difficulty of applying this approach is the precise collection of historical data for threat events probability calculation, which requires data input from those to be assessed Cloud Computing platforms and their vendors. Risk assessment during service construction, deployment, operation, and during admission control and internal operations is virtually nonexistent. There is a lack of structured analysis approaches that can be used for risk analysis in cloud computing environments. This framework doesn't cover risks during all the stages of the lifecycle of the service when it exists on the cloud [27].
[15]	Cloud providers and cloud consumers	There is a lack of structured analysis approaches that can be used for risk analysis in cloud computing environments.
[17]	Cloud environment	This work has not implemented such a service but rather offer it as a paradigm to be pursued. Automating risk assessment for cloud computing is far from lights to be established, because the risk assessment needs always judgments of experts to succeed
[18]	Cloud providers and cloud consumers	This framework can be used just to compare between cloud providers to select the best one basing on calculation of risk factor of each one
EBIOS [8] MEHARI [9] OCTAVE [10]		These methods don't include the specific characteristics of cloud computing Using these methods needs more time and more money due to the complex nature of cloud computing These methods are potentially cumbersome and contain several steps to validate
[16]	Cloud providers	There is a lack of complete model or method of risk assessment in cloud computing environment

From this study of current risk assessment for cloud computing, it is clear that at present there is a lack of risk assessment approaches for cloud consumers. A proper risk assessment approach will be of great help to both the service providers and the cloud consumers. With such an approach, the cloud consumers can check the effectiveness of the current security controls that protect an organization's assets and the service providers can maximize and win the trust of their cloud consumers if the level of risk is not high. Also the cloud consumers can perform the risk assessment to be aware of the risks and vulnerabilities present in the current cloud computing.

VI. CONCLUSION AND FUTUR WORK

After survey the literature of risk assessment regarding cloud computing, most of the current works is for helping cloud consumers assessing their risk before putting their critical data in a security sensitive cloud. Therefore, the most obvious finding to emerge from this study is that, there is a need of specific risk assessment approach. At present, there is a lack of structured method that can be used for risk assessment regarding cloud consumers to assess their resources putting outside in order to maximize the trust between the cloud consumers and cloud providers and also the effectiveness of the security system established.

As future work, we will develop a new risk assessment approach, which can take into account the complex nature of cloud computing environment.

REFERENCES

[1] Vivek Kundra. (2011, july) Seeking Alpha. [Online]. Available: http://seekingalpha.com/article/283444-cutting-government-spending-with-cloud-computing

[2] Rehan Saleem, "What's New About Cloud Computing Security?," 2011

[3] Mkpong-Ruffin, I., Umphress, D., Hamilton, J. and Gilbert, J. Quantitative software security risk assessment model , *ACM workshop on Quality of protection*, Alexandria, Virginia, USA, 2007.

[4] Mell P, Grance T. Perspectives on cloud computing and standards. National Institute of Standards and Technology (NIST). Information Technology Laboratory; 2009.

[5] CSS, White paper on software and service architectures, Infrastructures and Engineering – Action Paper on the area for the future EU competitiveness Volume 2: Background information, Version 1.3, retrieved:15.08.2010,http://www.euecss.eu/contents/documentation/volume%20two_ECSS%20White%20Paper.pdf

[6] R. Farrell, "Securing the cloud-governance, risk and compliance issues reign supreme," *Information Security Journal: A Global Perspective*, vol. 19, pp. 310–319, 2010.

[7] ISO 31000:2009, Risk management—Principles and guidelines

[8] EBIOS, Central Directorate for Information Systems Security, Version 2010 website. [Online]. Available: http://www.ssi.gouv.fr.

[9] Operationally Critical Threat, Asset and Vulnerability Evaluation (OCTAVE), Carnegie Mellon - Software Engineering Institute, Juin 1999.

[10] Method Harmonized Risk Analysis (MEHARI) Principles and mechanisms CLUSIF, Issue 3, October 2004.

[11] Catteddu, D., Hogben, G.: ENISA Cloud Computing Risk Assessment. ENISA (2009)

[12] Catteddu, D., Hogben, G.: Cloud Computing Information Assurance Framework. ENISA (2009)

[13] P. Saripalli and B. Walters, QUIRC: A Quantitative Impact and Risk Assessment Framework for Cloud Security , In the Proceedings of the IEEE 3rd International Conference on Cloud Computing, pp. 280-288, 2010

[14] H. A. Linstone, The Delphi Method: Techniques and Applications. Addison-Wesley, 1975.

[15] Amit Sangroya, Saurabh Kumar, Jaideep Dhok, Vasudeva Varma, "Towards Analyzing Data Security Risks in Cloud Computing Environments", International Conference on Information Systems, Technology, and Management (ICISTM 2010), Bangkok, Thailand

[16] J. Oriol Fitó, Mario Ma_as and Jordi Guitart, Towards Business driven Risk Management for Cloud Computing, pp. 238-241, Proceedings of 2010 Int. Conf. on Network and Service Management

[17] Burton S. Kaliski Jr. and Wayne Pauley "Toward Risk Assessment as a Service in Cloud Environments," *EMC Corporation, Hopkinton, MA*, USA 2010

[18] M. Hale, and R. Gamble, "SecAgreement: Advancing Security Risk Calculations in Cloud Services," *8th IEEE World Congress on Services*, 2012.

[19] Heiser, J., Nicolett, M.: Assessing the Security Risks of Cloud Computing. Gartner (2008)

[20] Vishal Visintine, "An Introduction to Information Risk Assessment", GSEC Practical ,Version 1.4b, August 8, 2003

[21] Feather, M. and Cornford, S. Quantitative risk-based requirements reasoning. *Requirements Engineering, 8* (4),pp. 248-265.

[22] Cloud Security Alliance (CSA): Top threats to cloud computing, version 1.0. http://www.cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf (March 2010)

[23] Ekelhart, Fenz, Klemen and Weippl, Security Ontologies: Improving Quantitative Risk Analysis, (2007), 156a.

[24] DCSSI (2004). EBIOS – Expression of Needs and Identification of Security Objectives. http://www.ssi.gouv.fr/en/condence/ebiospresentation.html, France.

[25] ISO/IEC 27005 (2008). Information technology -Security techniques - Information security risk management. International Organization for Standardization, Geneva.

[26] ISO/IEC 27001 (2005). Information technology -Security techniques - Information security management systems - Requirements. International Organization for Standardization, Geneva.

[27] Afnan Ullah K, Manuel O, Mariam K, Ming J, Karim D."Security risks and their management in Cloud Computing". 2012 IEEE 4th International Conference on Cloud Computing Technology and Science

[28] Van Scoy, Roger L. Software Development Risk: Opportunity, Not Problem

[29] Deloitte. Executive Forum - Cloud Computing: risks, mitigation strategies, and the role of Internal Audit. Available: http://www.deloitte.com

[30] C. Pettey and B. Tudor. *Gartner says worldwide cloud services market to surpass $68 billion in 2010* Available: http://www.gartner.com/it/page.jsp?id=1389313

[31] Press Office. (2010, 31 August 2010). *Cloud Computing Services - New Market Report Published*. Available: http://www.companiesandmarkets.com/r.ashx?id=41AETZYHJ289173&prk=ecb8413c602cb89051067456b636c7b9

[32] Peiyu L., Dong L., 2011. "The New risk assessment model for information system in Cloud Computing environment", Procedia Engineering 15, pp. 3200 – 3204

[33] NIST, "Standards for Security Categorization of Federal Information and Information Systems. *FIPS-199*," <csrc.nist.gov/publications/fips/fips199/FIPS-PUB-199.pdf>, Accessed Dec 2010.

[34] Harms-Ringdahl, L. (2001) Safety analysis: Principles and practice in occupational safety. CRC Press.

[35] Butler, S.A., Security Attribute Evaluation Method: A Cost-Benefit Approach, (2002), 232.

[36] Z. Xuan, N. Wuwong , et al., "Information security risk management framework for the Cloud Computing environments," in 2010 IEEE 10th International Conference on Computer and Information Technology (CIT), 2010, pp. 1328-1334.

[37] Gamal A. Awad, Elrasheed I. Sultan Noraziah Ahmad, N. Ithnan, "Multi-objectives model to process security risk assessment based on AHP-PSO" ,Modern Applied Science Vol. 5, No. 3; June 2011

[38] Theoharidou, M., Mylonas, A., Gritzalis, D.: A risk assessment method for smartphones. In: Proc. of 27th IFIP Information Security and Privacy Conference, pp. 428-440 (2012)

[39] X. Zhang, N. Wuwong, H. Li and X. Zhang, Information security risk management framework for the Cloud Computing environments,pp. 1328-1334,Proceedings of the 10th IEEE Int. Conference on Computer and Information Technology, 2010

[40] A. Morali and R. J. Wieringa, Risk-based confidentiality requirements specification for outsourced IT systems, pp. 199-208, Proceedings of the 18th IEEE International Requirements Engineering Conference, 2010, DOI 10.1109/RE.2010.30

[41] C. S. Yeo and R. Buyya, Integrated risk analysis for a commercial computing service in utility Computing, Journal of Grid Computing, Vol 7,No.1,pp.1-24,ISSN:1570-7873,Springer,Germany,March 2009

[42] Min Luo, Liang-Jie Zhang and Fengyun Lei, An Insurance Model for Guaranteeing Service Assurance, Integrity and QoS in Cloud Computing, pp. 584-591, Proceedings of 2010 IEEE International Conference on Web Services, DOI 10.1109/ICWS.2010.113

[43] A. Juan Ferrer, F. Hernandez, J. Tordsson, E. Elmroth, C. Zsigri, R. Sirvent, J. Guitart, R.M. Badia, K. Djemame, W. Ziegler, T. Dimitrakos, S.K. Nair, G. Kousiouris, K. Konstanteli, T. Varvarigou, B. Hudzia, A. Kipp, S. Wesner, M. Corrales, N. Forgo, T. Sharif, and C. Sheridan, OPTIMIS: a Holistic Approach to Cloud Service Provisioning, Future Generation Computer Systems, 2011,DOI: 10.1016/j.future.2011.05.022

[44] K. Djemame, I. Gourlay, J. Padgett, K. Voss, and O. Kao, Risk management in Grids, In R. Buyya and K. Bubendorfer, eds, Market-Oriented Grid and Utility Computing, pp. 335–353. Wiley, 2009

[45] J. A. Zachman, A Framework for information systems architecture,IBM Systems Journal, Vol 26. No 3, 1987

A Model of an E-Learning Web Site for Teaching and Evaluating Online.

Mohammed A. Amasha
College of Science and Arts, Computer Science Department, Qassim University
Alrass City, Saudi Arabia –Domyat University,Egypt

Salem Alkhalaf
College of Science and Arts, Computer Science Department, Qassim University
Alrass City, Saudi Arabia

***Abstract*—This research is endeavoring to design an e-learning web site on the internet having the course name as "Object Oriented Programming" (OOP) for the students of level four at Computer Science Department (CSD). This course is to be taught online (through web) and then a programme is to be designed to evaluate students' performance electronically while introducing a comparison between online teaching , e-evaluation and traditional methods of evaluation. The research seeks to lay out a futuristic perception that how the future online teaching and e-electronic evaluation should be the matter which highlights the importance of this research.**

Keywords—Key Words; e-learning; evaluation; designing a website; university performance

I. INTRODUCTION

The twenty-first century has witnessed the transformation of a new world in which contemporary information technology prevails in global society; where the educated people have a free access to the information they need within a short time with least efforts-the matter that results in enhancing the competence level of creativity and production [4].

The issue of educational development has topped the agenda of several countries on their conviction that the progress and promotion of their people begin with development of education. Hence, there are certain countries which have begun to amend their educational systems in their staunch faith that education is the only way to rise up to the developments and challenges characterizing the present era. These countries have sought to introduce modern technology to their educational system representing computer technologies in their bid to benefit from the potentials of computer as an educational aid and technological coordinator [20].

Still, a plethora of programmers and programming languages that aim at best optimizing the web in the domains of e-mails, e-conferences, and written/ voice-chat which have kept up with the emergence of the internet and its development. Thus, introducing the internet into education that has been given due care to help give students quick and easy access to knowledge [15].

The internet has remarkably and considerably contributed to the development of all aspects of the educational process. It doesn't only have a great bearing on educational curricula and methodology but it also extends to administrative domains. In addition, it helps to give a boost to the modern educational types as open education, distance education, self-learning and individual learning [17]. The benefits that internet can render are countless as it may bring about a genuine and effective development in Arab Educational Systems [15].

Educational development starts after accomplishing integrated procedures shaped into the development of a system, the foremost of which is giving due care to amending, improving educational curricula and methodology. Universities should be pioneers in developmental process as they are considered the locomotives of advancement and modernization in a society. Their message is to steer the process of development in the different institutions scientific knowledge and modern technology.

Due to this inspiration the educational process began to develop in all its spheres. In this context, some educational centres and units were established aiming at promoting the instructional, educational and research processes in order to achieve the goals of the ongoing development in an age of informational and developmental change. One of these units is E-learning Unit which has been playing the greatest role of developmental process.

This unit aims at the developing of the educational curricula in all specializations of the different colleges of the university in its pursuit to enhance the role of education in the realization of outputs of educational process. E-education helps to use strategies of competence and feedback in recognizing students' individual needs and devising ways to develop them and it also contributes to provide an intellectual forum, which aims to enhance the educational content and give a boost to the educational curricula. [1].

The current research attempts to design an an e-learning web site on the internet as a course namely, "Object Oriented Programming" (OOP) for the students of 4th level at CSD. OOP is a programming paradigm that represents concepts as "objects" that have data fields (attributes that describe the object) and associated procedures known as methods. Objects, which are instances of classes, are used to interact with one another to design applications and computer programs (www.wikipadeia.com).

This course is to be taught electronically (via the web) and then a programme is to be designed to e-valuate students and draw a comparison between e-evaluation and the traditional methods of evaluation. The research seeks to put forward a futuristic perception to the way future e-evaluation should be, the matter that highlights the importance of the research.

II. Statement of the Problem

The previous presentation demonstrates the importance of using E-learning in university education, the point that has been sustained by several studies carried out in this field, which stresses on the importance of disseminating the culture of education amongst the teachers and their assistants, developing university e- courses and making the students best optimize these courses.

These studies highlight the importance of transforming evaluation methods to cope with the development that's taking place. Thus, change of test patterns has become a necessity as it is one of the characteristics of this age and a prerequisite to keep abreast with the successive developments of today's world. In addition, evaluation is considered a substantial element in the educational curriculum and it's, therefore, one of the foundations of education development because any development in the aims, content and methodologies of a given educational curriculum that cannot be effected without reliance on evaluation results. It's a must to introduce new techniques in the bases and sources of evaluation via modern technology in order to promote and develop tools of evaluation.

There upon, employing that e-evaluation has become a pressing need as it guarantees the credibility to scores and it secures the grader to be unprejudiced. Besides, e-evaluation helps to solve the problems of the staff represented in giving tests and, grading them, assessing students' performance especially in the light of the appalling number of students overcrowding the colleges of the university at present. Hence, the importance of e-evaluation lies. The question can now be phrased as follows:

Can e-evaluation be effective and fair with students? Can we safely state that the time is now ripe to manage without the traditional (paper) evaluation and to resort to e-evaluation? Is there an effective vision to the future form of e-evaluation? And do students have a real interest in using E-evaluation? All these significant points need to be made clear. The current research attempts to make a comparison between e-evaluation and the traditional methods of evaluation and to survey the students' opinions in connection with e-evaluation & employing the students' opinions poll to form a proposed future vision of e-evaluation.

III. The Research Problem can be couched in the Following Questions:

- Do e-evaluation techniques result in an improvement on the attainment level of the sample students?
- Does the use of e-evaluation method takes less time than the conventional method?
- What do the students' think of the proposed programme using an e-evaluation technique?
- What's the future vision of the e-evaluation programme should be taken on?

IV. Research Aims

- To design a model for an e-course on the internet on object-based programming for the level four students, Computer Science Department (CSD).
- To Plan an e-programme on the course of object based programming to assess the performance of CSD electronically via the web.
- To identify which method works well with the students' polls, e-evaluation method or the traditional one.
- To investigate the students' viewpoints on e-evaluation method applied in this research to e-evaluate the students.
- To put a form of future vision of the e- evaluation to be acquired.

V. Significance of the Research

- Putting forward a model for an e-course via the internet beneficial to working out more e-courses in higher education.
- Endeavoring to draw up a future form to an e-evaluation programme that can contribute to eliminate difficulties associated with the traditional evaluation method represented in handing and collecting exam papers and estimating them in order to make best use of the teacher and learner's time.
- Studying the feasibility of the idea, analyzing its application and employing it to draw some conclusions in an attempt to contribute to the improvement and development of the educational process and to make best use of computer and the internet technologies in opening up new horizons and prospects in this field.

VI. Research Methods

The researcher will use the experimental method as it cares to study factors and variables that affect the phenomena or problem and changes in some of its aspects with other stables to reach the causative relationship between these variables.

VII. Research Sample

The sample of the research consists of 36 students picked out randomly from the level four students of CSD, at Alrass Faculty of Science and Art. The sample is divided at random into three groups; two experimental groups and one control group as follows:

The first Experimental Group: consists of (12) male students who got their learning through e-course technique and have been e-evaluated.

The second Experimental Group: is comprising of (12) male students that received their learning via e-learning but evaluated by the traditional evaluation method, i.e., paper examination.

The control Group: includes (12) students of both sex who have learned according to the traditional education and have been evaluated via the traditional evaluation method, i.e. paper examination.

VIII. Research Outline

This research is restricted to the second unit in Object Oriented Programming (OOP) course, which is a Programme using Action Script (PAS) being taught to the students of 4th level CSD, faculty of Science and Arts, Alrass.

- Assumptions
- Using e-tests to evaluate the students' cognitive attainment is better than using the traditional evaluation method (i.e, paper tests) and there is a statistically significant distinction at a rate of 0.05 of the arithmetic mean between the degrees of the three groups (the two experimental groups and the control one) in favour of the first experimental group.
- There is a statistically significant distinction at a level of 0.05 of the arithmetic mean between the degrees of the three groups (the two experimental groups and the control one) in the time of carrying out the test in favour of the first experimental group, which adopted e-evaluation method.
- The students' attitudes are positive towards e-learning using e-course and e-evaluation method followed in assessing the attainment level of the sample students.

IX. Research Terminology

A. The Internet

It's a technology linked with millions of computers connecting and allowing people together around the world to exchange information and ideas (proven E, 1999)

It's a multitude of computers communicating with one another; where millions of computers exchange information via the multiplex World Wide Web (Farhan N, Rafiq D., 1998)

Procedure

B. Planning the e-course and uploading it on the web

The researcher planned an e-course for the subject of OOP planned for the students of level four at computer science department, Alrass Faculty of Science and Arts in accordance with the criteria specified by E-learning Unit at the University for Planning E-courses. The e-course has been uploaded on the website of the following site:

http://mansvu.mans.edu.eg/moodle//mod/resource/view.php?id=2936

C. The course is designed according to the substantial procedures and steps followed in this respect as follows:

Pinpointing the learning topic: The second unit in OOP course is entitled as "Programme using Action Script" (PAS) designed for the students of fourth level at computer science department.

D. Analyzing the Subject Matter

The content is compiled according to the nature of the subject of which the topic of the course is selected. The researcher turned to the content of the university book relevant to the course and a simulation of the programme has been prepared in the form of an educational booklet before planning it into an e-programme.

X. Identifying the Educational Aims of the Programme

The educational aims of the programme fall into three categories:

- Cognitive aims: that is concerned with information and facts.
- Psychomotor aims: that tackles manual skills.
- Emotional aims: that deals with attitudes and values

XI. Writing the Subject Matter of the Programme

The content of the course has been drafted including a home page that will be displayed when the course website is uploaded (via the University Website). The page includes a number of entries to different parts of the course. The subject matter has been turned to jury to give their opinion about it before embarking on the e-programming process.

XII. E-Programming

The researcher employs the following pogrammes and languages to prepare the course:

- Sound Forge – Front page - Action Script- Switch MAX – flash CS2.

Evaluating the course and Turning it to the concerned experts:

- After preparing the course, the researcher referred it to a group of experts from university professors specialized in computer sciences to give their opinions on the feasibility of the programme.

XIII. Planning the Cognitive Attainment Test

The aim has been pinpointed, i.e., assessing the sample students' attainment of the concepts encompassed in the e-course which is prepared and edited on the internet through course teaching.

Specifications and Timetables have been prepared according to the compiled information which involves the topics that the test should include learning results which should be tested in accordance with the educational aims and the proportional significance of the subjects (proportional weight), [5]. Test specification table has been prescribed as follows:

Determining the number of expected in the test questions:

TABLE :SHOWS THE PROPORTIONAL WEIGHTS OF ATTAINMENT TEST ON SAP:

NO	Content	Aims level			Total of content weights
		application	understanding	cognition	
1	Writing action code and the types of data.	2	12	6	20%
2	Orders of Movie Control.	3	18	9	30%
3	Orders of mathematical function.	3	18	9	30%
4	Working with properties.	2	12	6	2%
	Total weights of the aims.	30%	60%	10%	100%

Identifying the Proportional Significance and Aims (Table of Aims)

TABLE SHOWS THE NUMBER OF QUESTIONS IN THE ATTAINMENT TEST ON (PAS):

NO	Content	Aims level			Total of content weights
		application	understanding	cognition	
1	Writing action code and the types of data.	0	1	1	2
2	Orders of Movie Control.	1	3	2	3
3	Orders of mathematical function.	1	3	2	3
4	Working with properties.	1	2	1	2
	Total weights of the aims.				10

- Phrasing the questions of the test in such a way that makes these questions easily-understandable and accurately specified.
- Giving instructions in the beginning of the test that show the aim underlying it and define questions and how to answer them. A simulation has been carried out and the validity of the test has been checked through measuring the validity of its subject matter by turning it to a number of specialists from the teaching staff majored in the field of specialization in this respect.
- Ensuring the stability of the test that has been given to a pilot population composed of 10 individuals with a 20 day interval between the first and the second test. Using Richardson' equation, stability co-efficient is found to be (0.86), an evidence to the stability of the test.
- Time duration of the test has been worked out according to the following equation:
- Time duration of the test = (The time the fastest student took + the time the slowest students took) /2.After working out time duration, the validity and stability co-efficient, thus, validates the test which is composed of 10 questions carrying two marks each.

XIV. DESIGNING THE TEST ON COMPUTER

To design the test in its final form the researcher followed the following steps:

- Eight models of the test have been designed on computer making a password for each and a username. Each student logs unto the home page of the test where he / she is asked to click a button which generates a number of tests at random, where he/ she takes one of them at random. At the end, the mark and the percentage are given in a window prepared for this purpose.
- As for the students who are given their tests according to the traditional evaluation method (paper tests), each was asked to pick a number randomly from a set of numbers and then he / she answer the random test. The test is corrected according to the correction key prepared for this purpose.
- Preparing a questionnaire based on the students' attitudes towards e-evaluation: This opinion poll aims at recognizing the viewpoints of the fourth level students sample consisting of 25 students of both sex at Computer Department, faculty of science and art as for their attitudes towards e-evaluation method. The questionnaire includes 20 phrases in connection with the test's range of accuracy its significance in assessing what it's built for, the effectiveness of the test's procedures, and its role in saving time and effort for each of the examiner and the examinee.
- The questionnaire has been referred to a group of experts in order to give their opinion with regard to its programme phrasing. The researcher worked out the percentage of the experts' approval of each phrase ruling out the points which didn't receive 85% at least of the jury's approval.

XV. DISPLAYING RESEARCH RESULTS

A. *The 1st Assumption*

Using an e-test to evaluate the students' cognitive attainment is better than the traditional evaluation method (i.e paper test). There is a statistically significant distinction at a rate of 0.05 between the arithmetical mean of the degrees of the three groups in favour of the first experimental group.

- To test the validity of this assumption, the value of (X2) for the statistical significance between the arithmetical means of the two experimental groups and the control group has been worked out by using "kruskal wallis" co-efficient for three independent and identical samples

of the cognitive attainment test degrees. The following table shows it as follows:

- The value of (X2) for the statistical significance of the arithmetical means of the degrees of the two experimental groups and the control one in the cognitive attainment test.

TABLE NO (3): SHOWS THE VALUE OF X2 FOR THE TWO EXPERIMENTAL GROUPS AND THE (CONTROL ONE) IN THE COGNITIVE TEST

Group	Rank of means	(fd)	kruskal Wallis value (X2)	Sig.
The 1st experimental group	29.54	2	19.45	A statistical significance at a rate of 0.000 in favour of the 1st experimental group
The 2nd experimental group	13.63			
The Control group	12.33			

Results indicate that there is a statistically significant distinction between the arithmetical mean of the students' degrees of the two experimental groups and the control one in favour of the first experimental group students' degrees. The distinction is essential and it doesn't trace back to the work of chance as kruskal Wallis' value of (X2) hit (19.945) to be bigger than the value of (X2) which struck (2.7) of rate of 0.000 and a freedom degree of 2.

It's noticed that mean of the attainment degrees of the second experimental group $\overline{X}$ equal (13.63) occurs after the first experimental group while the control group comes at last as for the mean of the test degrees $\overline{X}$ equal (12.33) and this is because the first and the second experimental groups received e-learning programme.

The preponderance of the first experimental group over the second experimental one in the test is due to the use of e-evaluation which helps students to remember and retrieve the information easily. Furthermore, the programme motivated them to improve their performance in the following question, the feeling that fosters examinee's self-confidence that helps him/her get better marks. The examine, moreover, transfers the experience he/she got while dealing with the e-course to the e-test he/ she takes finding no difference between what they learnt and what they have been evaluated in, the matter that the traditional evaluation method students didn't boast as they didn't have the opportunity to learn via a prepared e-course programme.

B. The 2nd Assumption

There is a statistically significant distinction at a level of 0.05 between the arithmetical mean of the degrees of the three groups(the two experimental groups and the control one) in the test's time duration in favour of the 1st experimental group which uses e-test in evaluation.

To test the validity of that assumption, the (x2) value of the distinction between the arithmetical means of the two experimental groups and the control one has been calculated by using kruskal Wallis co-efficient of their independent samples. The following table shows this as follows.

TABLE NO (3) (X2) VALUE (KRUSKAL WALLIS) OF THE DISTINCTION BETWEEN THE MEANS OF THE DEGREES OF THE TWO GROUP (THE TWO EXPERIMENTAL GROUPS AND THE CONTROL ON) ABOUT ATTAINMENT TEST'S TIME DURATION:

Group	Rank of means	df	kruskal wallis value (X2)	Sig.
The 1st experimental group	10.46	2	19.517	A statistical significance at a rate of 0.000 in favour of the 1st experimental group
The 2nd experimental group	16.08			
Control group	28.96			

N1= N2 =N3 = 12

Results indicate that there's a statistically significant distinction between the arithmetical mean of degrees of the two experimental groups and the control one in favour of the 1st experimental group, and that this distinction is not attributed to coincidence as the calculated kruskal's value of X2 hit (19.52) bigger than x2 value which reached (2.7) at a level of 0.000 and freedom degrees 2.

This points out that the first experimental group which used e-test in evaluating students' cognitive attainment, finished their test in time duration less than the other two groups. This is due to flexibility of dealing with the test and the clarity of its instructions. Moreover, providing the test with a direct feedback to the answers helps students move quickly from one question to the following and vice versa, the matter that results in concluding the test more quickly if compared to the traditional paper tests which fall short of achieving it.

C. The 3rd Assumption

"The students attitudes towards e-learning using e-courses and e-tests designed to assess the cognitive attainment level of the student sample are positive"

XVI. THE VALIDITY OF THE ASSUMPTION

To test the validity of this assumptions, the researcher, uses the general method of calculating X2 of the frequency table 1 ×2 to work out the statistical significance of frequency distinctions between the students approvals and disapprovals of each phrase of the questionnaire around e-evaluation method used in evaluating cognitive attainment. The following table shows this as follows:

TABLE NO (4) SHOWS FREQUENCIES, PERCENTAGE, X2 VALUE, AND THE STATISTICAL SIGNIFICANCE OF STUDENTS' RESPONSES TO THE FIRST PIVOT CONCERNED WITH "THE EFFECTIVENESS OF THE PROGRAMME USED IN EVALUATION".

NO	Phrases	agree		disagree		X2
		f	%	F	%	
1	The e-test assess the students' cognitive attainment more accurately than the paper one	19	76%	6	24%	6.76
2	The test highly	21	84%	4	16%	11.56

	depends on accuracy					
3	The method of the test is more effective efficient and it's time and effort saving for both the examiner and the examinee than paper tests	20	80%	5	20%	9
4	The questions are couched in such a clear and simple way that's apprehensible to the learner	18	72%	7	28%	4.84
5	The test is comprehensive; covering all parts of the course.	22	88%	3	12%	14.44
6	I feel that the answer technique followed in the test is up-to-date.	24	96%	1	4%	21.16
7	The method of the test copes with e-methodology.	22	88%	3	12%	14.44

The pervious table shows that x2 value of all phrases concerned with the students' view points on the first pivot of the questionnaire that deals with "the effectiveness of the programme used in evaluation " is bigger than the x2 value which struck 3.84 at a statistically significant rate of 0.05 and (1) freedom degree.

This indicates that the distinction between the observed frequencies and the expected ones around the phrases concerned with that pivot is statistically significant and it doesn't, therefore, go back to chance factor. The distinction between the students' approvals or disapprovals to the phrases of the first pivot in the questionnaire is in favour of the use of e-evaluation method.

On other words, the students agreed that e-evaluation method is effective, efficient and saving effort and time for the examinee and that the style of e-evaluation copes with the style of e-learning followed in the e-course. They also described the test as a developed one that doesn't allow cheating. In addition, there is a general satisfaction among the students due to the non-intervention of the human element in evaluation process. Thus, students feel completely satisfied with the result. They maintained that the test is in general better than paper test method used in the past. The student, furthermore, strongly stressed the importance of using e-evaluation method in all tests not only in the test of the student's performance marks of the year applied in this study.

TABLE NO (5) SHOWS THE FREQUENCIES, PERCENTAGES AND THE VALUE OF X2 AND ITS STATISTICAL SIGNIFICANCE OF THE STUDENTS' RESPONSE TO THE FIRST PIVOT CONCERNED WITH THE VALIDITY OF E-TESTING.

NO	Phrases	Agree		Disagree		X2
		f	%	F	%	
1	The test offers the opportunities to try the wrong answers again.	20	80%	5	20%	9
2	I think that multiple correct answers fares well during the test.	7	28%	18	72%	4.84
3	I prefer the technique of putting out the result immediately.	18	72%	7	28%	4.84
4	The style of the test is developed and bars cheating.	24	94%	1	4%	21.16
5	I feel that the non-intervention of the human element in evaluation process is much better.	18	72%	7	28%	4.84
6	I think that the possibility of correcting the answers helped me improve my score	18	72%	7	28%	4.84
7	I feel stress and fear during carrying out my test.	8	32%	17	68%	3.24
8	Reinforcement is direct, exciting and untraditional.	19	76%	6	24%	6.76
9	I'm satisfied with getting my score without the intervention of the examiner	23	92%	2	8%	17.64
10	Using objective question in the test is much more better than open-ended questions followed in paper test	21	84%	4	16%	11.56
11	I think that e-evaluation method should be used in all tests not only in student performance score of the year.	25	100%	-	0%	25

The previous table shows that the value of x2 of all the phrase concerned with the students' responses on the second dimension of the questionnaire that deals with "the validity of e-testing" is bigger than the value of x2 which struck 3.84 at a statistically significant rate of 0.05 and a freedom degree of 1. This indicates that the distinction between the observed frequencies and the expected ones on the phrases in respect is statistically significant and not attributed to coincidence factor.

The results indicated that the students agreed that the questions of the test had been phrased in a clear and simple way and that the test covers all parts of the course. The students also think that direct evaluation and giving them the opportunity to correct their answers helped them improve their scores, procedures that are not used in the traditional evaluation method (i.e, paper test).

Some students objected to giving more than one correct answer in the tests and complained that the instructions of the test didn't refer to this. Students also pointed out that direct evaluation of the questions made them feel excited, suspense and lack of traditionalism. The students, in addition, stated their un satisfaction with fill in the space questions because e-evaluating these questions may be inaccurate due to the difficulty of finding the typical answer. They also adopt the same attitude towards open-ended question in the questionnaire that deals with their viewpoints in general.

In connection to the open-ended question, the students commented in phrases which were repeated in a few sheets as follows:

- The test depends highly on accuracy and this is the best of its characteristics.
- The best thing in this test is putting out the result immediately. (Test score)
- The best thing in this test is that the test and the questions are picked out randomly, the matter that decreases cheating.
- The sole obvious shortcoming in this test is that it concentrated on to one question type.
- The method of the test is developed well
- There is more than one correct answer to some questions but it's is preferred to have just one answer for the question.
- The possibility of correcting question answer and trying it again is considered one good feature of that test(the possibility of moving from one question to another and vice versa)
- "Help" should be annexed to home page to define the properties of the test and the buttons
- The test should be limited to definite time duration.
- This method is progressive and up-to date and it doesn't depend on the traditional method of the previous tests.
- It's preferred to train students on the test and the endorsement of the score should be related to time.

XVII. The Proposed Future Vision of an E-Test

- Introduction of the test and presentation technique:
- Designing the introduction in such an attractive way that attracts the learner's attention and galvanises his motivation.
- The introduction includes an emulation model putting forth the way of answering the test and a window is made for user data.
- Moving to a following page which includes the examinee's data (username- password) linked to the student's database.
- Linking the test with Question Bank) database with a button that generates the number of test questions randomly from the database according to the number of questions the examinee will determine.
- Moving to the start page of the test where a random test is generated according to the number of questions set before.
- Planning a feedback to give the learner the opportunity to try again incorrect answers.

XVIII. Data of Final Test Results

- After concluding question answers a button is activated to send the student's result to his/ her database where the programme puts out his / her score in the designed place.
- The test is administered according to general time duration not to specify time for each part of the test where the programme is immediately shut down after the time is over.
- It's available to the learner to print a report.

XIX. Conclusion

In the light of applying research procedures to the current population, the following conclusion can be drawn out:

- Ensuring the availability of electronic learning for the students via the internet and by the educational banks help galvanize / give an impetus to students' motivation and enthusiasm for learning.
- The use of e-evaluation provides students with a sense of satisfaction and helps them greatly to get better score as it's characterized by accuracy, objectivity and its use of simulation technique which helps students easily retrieve information.
- The students have a high opinion on learning style using e-course via the internet and also towards e-evaluation method.
- It's strongly recommended to train students to use the internet and how to deal with e-courses and e-tests before using and applying this type of tests.
- The use of modern technology in education results in a significant development in the educational process as well as in students' thoughts and attitudes. In other words, the use of modern technology contributes to improving the educational product and enhancing quality of the learner.
- The provision of an electronic means to evaluate the students makes it easier for the staffs do their duties and give them plenty of time to unleash their creativity and achievement in their field of specialization.
- Planning e-tests, e-grading and writing down the marks get rid the staff and their assistants of the concomitant difficulties concerning handing over and collecting exam sheets followed in the traditional evaluation method (i.e, paper test).
- The use of e-test results in saving the time of both the examiner and the examinees, the matter that allows them make maximum use of their time in other business.

XX. Recommendations

- Generalizing the use of e-evaluation method to high education courses via the internet.
- Disseminating the culture of e-learning and e-evaluation amongst the staff and their assistants through holding training sessions and periodicals.

- Holding training sessions with the staff and their assistants around how to prepare e-questions, e-tests and how to design a site for them on the web.
- Training undergraduate students on how to deal with the internet in order to be able to deal with e-learning and e-evaluation.
- Expansion in building educational banks for high education curricula and the encouragement of designing e-course on the web.
- Studying the feasibility of the idea, analyzing its application, and benefiting of the proposed e-design in this study to plan similar models that may enhance e-evaluation.
- Establishing a centre for e-evaluation that includes a number of specialists in assessment, evaluation and curricula in addition to computer and education technology teachers concerned with developing the ongoing evaluation process inside the university and with setting up and promoting e-question banks in order to give the student access to continuous training on that type of questions via the internet.

REFERENCES

[1] Abdel Hadi M. "E-learning via the Internet", Cairo, the Lebanese Publishing House, 2005, P.106

[2] Abdel-Malik A. "Supervision and Evaluation Service of Community", Cairo, AL-Anglo Library, 1976, P. 71

[3] Al-Sunbul A. "Justification of Adopting Distance Education in the Arab World", Education Bulletin, the Qatari National Committee for Education, Culture and Sciences, Issue No.137, 30th year, June, 2001

[4] Saddiq M "e-publishing in Balance", the Education Bulletin, the Qatari National Committee for Education, Culture and Science, Issue No. 144, March 2003, p. 60

[5] Hattab F ,Etal. "Psychological Evaluation", Al-Anglo al-Masriyya library, Cairo, 1987, p.397

[6] Himdan M. The role of Modern computer technology and the Internet in the vocational Development of High School Staff", the Education Bulletin, the Qatari National Committee for Education, Culture and Sciences, Issue No. 146, 32ndyear, Sep. , 2003, PP. 242-261

[7] Hung D. "Constructivism and e-learning: balancing between the individual and social levels of cognition", educational technology, vol.xli, no., 2, march-april, 2001.

[8] Ibrahim M. "Future Visions in Modernizing the Educational System", Cairo, al Anglo al-Masriyya Library, 2001, pp. 217-225

[9] Kamel N. "The internet and the Globalization of Education and its Development", the Education Bulletin, Issue No.303, 29th year, July2000, pp. 349-360

[10] Lyman M. Ettal "Assessing assessment: the inequality of electronic testing, webaim, jun 2004, [online] www.webaim.org/coordination/articles/assessment.

[11] Lyman M. Ettal "Assessing assessment: the inequality of electronic testing, webaim, jun 2004, [online] www.webaim.org/coordination/articles/assessment.

[12] M.Mwale H. "Safe school for teaching and learning: developing a school-wide, self-study process", Blacksburg, Virginia, 2006.

[13] Mahran M., Et al. "Designing "Preprogrammed In Arabic and English to Manage Distance Learning Via The Internet", A published Study, Centre of Researchers and Development, Ministry of Defence-cairo, 2005.

[14] Muhammed F. "E-Book and The Development of Children's thinking Abilities", Education Bulletin, the National Committee for Education, Culture and Science, Issue No. 136, Sep.,2003, p. 276

[15] Nadahi H. "Use and documentation of electronic information: a survey of eastern regional technology education collegiate association students", journal of technology education, vol.14, no.2, spring 2003.

[16] Peggy m.n and david w.l. "Activity theory as a framework for analyzing CBT and e-learning environments", educational technology, vol.xli, no.4, july-agust 2001.

[17] Schiftre C. "Faculty motivators and inhibitors for participation in distance education" education technologies, vol.xl, no.2, march-april, 2000.

[18] Schiftre C. "Faculty motivators and inhibitors for participation in distance education" education technologies, vol.xl, no.2, march-april, 2000.

[19] Shihata H, Et al. "High School Teaching and Evaluation Future Critique" Arab Gulf Massage, Issue No.78, 21st year, 2001, pp.13-50

[20] Betty collies "The Internet as an Educational Innovation: Lessons From Experience with Computer in Plementation", Educational Technology, vol.xxxi, No.11, November.

The Failure of E-government in Jordan to Fulfill Potential

Raed Kanaan
Faculty of Informatics
Amman Arab University
Amman, Jordan

Ghassan Kanaan
Faculty of Informatics
Amman Arab University
Amman, Jordan

***Abstract*—The aim of this paper is to uncover the reasons behind what so-called total failure in e-government project in Jordan. Reviewing the published papers in this context revealed that both citizens and employees do not understand the current status of this program. The majority of these papers measure the quality of e-services presented by e-government. However, according to the minister of Communication and Information Technologies (MOCIT), only three e-services are provided by this program up to writing this paper. Moreover, he decided to freeze the current working on e-government programme. These facts drove the authors to conduct this research. General review of the existing literature concerning e-government implementation in Jordan was applied, then a qualitative research was utilised to uncover the reasons behind the failure of the e-government program in Jordan. The collected data then was analysed using Strauss and Corbin's method of grounded theory. This paper illustrates that Jordanian government need to exert strenuous efforts to move from the first stage of e-government implementation into an interactive one after fourteen years of launching the program, considering that only three e-services are presented up to October 2013. Reasons behind the failure of e-government in Jordan have also been identified.**

Keywords—e-government; grounded theory; Jordan; total failure

I. INTRODUCTION

E-government has become a focus of government efforts in several countries around the world. Jordan was one of the first developing countries that started e-government implementation in the year of 2000.

E-government can be defined as "the carrying out of governmental activities using Information and Communication Technology tools in order to deliver better services to citizens, businesses and government entities (including government employees)" [1]. Therefore, e-government is expected to improve efficiency, transparency and effectiveness of the services provided by government departments to citizens, businesses and government itself [2] [3].

Several benefits and advantages of e-government have been listed repeatedly in the literature, table 1 summarises these benefits for both developed and developing countries from different perspectives:

TABLE I. BENEFITS OF E-GOVERNMENT IMPLEMENTATION (ADAPTED FROM [1])

Perspective	Benefits	Source
Government	Reducing errors Saving time and money Reducing bureaucracy Improving the quality of services Increasing economic competitiveness Increasing accountability Driver for other companies	[4][5][6][7][8]
Citizen and businesses	Services available 24/7 Increasing citizen participations in government decision making Bridging the digital divide Increasing transparency	[9][10][11][12][13]

In order to achieve the above mentioned benefits, several stages of the e-government implementation need to be followed in. There are different stage models in the literature as described in table 2:

TABLE II. E-GOVERNMENT STAGES (ADAPTED FROM [1])

Stages	Definition	Source
Publish	Presenting government information online to citizens.	[14]
Interact	Two way communication between government and citizens and their involvement in government processes.	
Transact	Conducting all transactions online.	

Cataloguing	Presenting government information online via web sites.	[15]
Transaction	Citizens interact with government electronically.	
Vertical integration	Local systems linked to higher level systems.	
Horizontal integration	Systems integrated across different functions.	
Information publishing	Each government department creates a web site.	[16]
Official two-way transactions	Citizens make electronic transactions such as paying tax and buying TV licenses.	
Multi-purpose portals	Creation of a single point (portal) to enable citizens to access and obtain government information and services.	
Portal personalisation	Citizens have the ability to customise portals according to their needs.	
Clustering of common services	Government departments will disappear when the portals become better.	
Full integration and enterprise transformation	Fully better changing in government departments.	
Presence	Presenting web sites and providing information about departments.	[17]
Interaction	Downloading electronic forms.	
Transaction	One-way communication.	
Transformation	Two-way communication, streamlining of procedures.	
Information	Government posts information on its web sites.	[18]
Two-way communication	Citizens communicate online with Government and they can fill in forms and request information or services.	
Transaction	All transactions conducted online.	
Integration	Citizens can access all services via single portal.	
Participation	Political participation such as voting online, and participating in decision making by posting comments and suggestions.	

However, researchers still argue that e-government has not yet reached the promise of the above mentioned benefits especially in developing countries. Reference [19] stated that barriers to the successful implementation of e-government are related to the failure of a government to accomplish fundamental requirements supporting the initiative of e-government. Moreover, [20] described e-government program in developing countries as a recipe for failure. He defined the failure as the "inability of such a system to achieve predefined goals or other, previously unanticipated benefits (p.2)". That could be the case in Jordan due to the fact that the initiative was never implemented or was implemented but immediately abandoned [21]. This paper therefore aims to provide an insight to the core reasons behind the failure of e-government project in Jordan since there are only three e-services presented by the program from 2000 to 2013, according to the minister of CIT.

II. Literature Review On Jordan And It's E-Government Program

Jordan is a relatively small country of 7,371,000 million people. It is located in the Middle East within a total area of 92,300 Square Kilometres, the vast majority of which is either desert or semi-arid. Jordan neighbouring countries include Iraq, Israel, Saudi Arabia, Syria and the West Bank [22], and the whole area can be described as "volatile".

As part of the king's strategy for economic growth, the Jordanian government embarked on a major long-term e-government initiative in the year 2000. The project aimed to deliver positive change to both government and its services, by improving service delivery, enhancing responsiveness to customer needs, increasing transparency (and thus reducing the potential for corruption) and efficiency of operations, and enhancing the level of understanding of ICT in general with Jordanian society [23]via the use of e-government.

However, thirteen years on and the e-government project in Jordan still falls predominantly within the informational and publishing stage of e-government evolution [18] [14] as described in the next paragraphs. For instance, [24] an e-government readiness assessment model was proposed basing on six factors. These factors are as follows: organizational readiness, governance and leadership readiness, customer readiness, competency readiness, technology readiness, and legal readiness. They studied each factor and resulted in some important issues that need to be considered before implementing e-government in Jordan. They claimed that "E-Government implementation will not face major barriers if Jordanian government adopt their model and/or take into consideration the unsolved problems in each of the above mentioned six factors.

Reference [25] studied the e-government implementation progress by evaluating the strength and weaknesses in selected government websites. They referred to the limitation in the online business as a result of the lack and absence of local certificate in Jordan. They concluded that private sector needs to get involved in building the required infrastructure of the e-government. In addition, they asserted that increasing the awareness of the customers on how to practice e-business is an important factor in successful e-government in Jordan. Furthermore, [26] when examined G2B practices in Jordan, they suggested that government should increase the number of internet users before moving toward more advanced level of e-government implementation. However, even if the internet users have been increased there are no available e-services to be used by the citizens or businesses. Nonetheless, [27] and [25] considered the e-services being provided by banks or the private sector as part of e-government services. Several researchers have confirmed that there are clear differences between e-government and e-commerce though. For instance, [28] they asserted that e-commerce customers are not the equivalent of citizens in e-government. Moreover, government acts as a representative of its citizens, while companies are in a competition with one another. Therefore, services offered by banks (i.e. private sector) cannot be considered as e-government services.

Reference [29] described the history of the e-government project in Jordan. They return the slow progress in e-government implementation in Jordan to the fact that the majority of the projects are not linked together or activated. From the authors' point of view, [29] utilised qualitative methodology in their research; however they did not mention how the collected data was analysed.

Reference [30] claimed that the use of e-government services will be increased by citizens if the websites is designed carefully. Furthermore, [30] suggested that designers of e-government services should allow the users to interact freely with the websites and this consequently will increase their intention to use e-government services.

To conclude, it seems to the researchers that the existing body of literature relating e-government in Jordan focused on issues such as intention to use e-government services, e-commerce transactions as part of e-government, narrative story of the development of e-government project in Jordan, the weaknesses of the ministries websites. However, none of the above mentioned literature addressed the reasons behind the failure of the project, hence the aim of this paper. The next section will discuss the qualitative methodology that will be utilised to provide more insight on the reasons of this failure.

III. Research Methodology

A qualitative method was applied in this research in order to gain rich understanding of such reasons. The employed data collection method was semi-structured interviews with public sector employees. According to [31], interviews of this nature of research tend to reach a point of data saturation after interviewing about eight individuals. Therefore, a total of eight employees were interviewed, these included three from MOICT and five from other ministries involved with the MOICT with the project. Questions were open in nature, allowing each interviewee to articulate his/her viewpoints as to why s/he felt the e-government programme can be categorised as total failure [21]. Secondary data, such as newspaper articles, were also utilised where there is relevant information to e-government programme.

Strauss and Corbin variant of grounded theory [32] was used as data analysis tool. This variant was chosen because of its recognition of the use and influence of existing technical literature, and the availability of strong guidance in its application [33]. Open coding was followed by axial coding which was followed by selective coding. Through this sequence several reasons behind the failure of e-government project in Jordan were identified; more discussion will be provided in the next section of this paper.

IV. Results And Discussion

The following factors have been identified as a core reasons behind the failure of e-government project in Jordan, these are as follows:

A. economic situation

From the grounded theory analysis, difficulties in the economic situation in Jordan emerged as a very strong reason behind the failure of e-government project.

The following are a sample of quotations from the interview transcripts that explicitly or implicitly refer to economic situation and its impact. An interviewee in the ministry of finance mentioned:

"Deficit in the government budget is always predominant, as an employee in the Ministry of Finance, I never remember that there is money surplus. So how such projects will be succeeded"

Lack of funding was a persistent barrier to the development in e-government project in developing countries [34]. However, in Jordan the issue is quite different because lack of funding represent constant problem, not only for e-government project, but also for other projects unlike ICT. The above interviewee added:

"Even if the money is available I don't think the e-government will be taken as high priority over any other important projects or initiatives"

Therefore, it is obvious that the failure in e-government implementation in Jordan may not be solved in the short run; this is due to the fact that lack of funding in the government budget is persistent for several years.

B. corruption

Corruption was also identified as a major reason behind the failure of e-government project in Jordan. All interviewees asserted that unless government fight corruption, none of the government project will be succeeded. They refer to the corruption as an important reason for not being a civilised country (i.e. Jordan). One interviewee claimed:

"You know how many millions have been stolen by officials even by security ones"

As [35] confirmed, Jordan score in the corruption perception index is 48 (Scores range from 0 (highly corrupt) to 100 (very clean)). Moreover, Jordan score in the control of corruption was 0.040348086 (Point estimates range from about -2.5 to 2.5. Higher values correspond to better governance outcomes) in 2010 which reflects perceptions of the extent to which public power is exercised for private gain. These statistics confirmed that corruption in Jordan is widely spread and practiced by public sector employees. Corruption is not only resulted in e-government failure, but also many projects have failed as a result of it.

Furthermore, two interviewees from ministry of Planning and International Cooperation, and ministry of Social Development said:

"The corruption is reached to e-government department itself. The salaries are very high in comparison to other salaries in other ministries"

Ironically, citizens expect from the e-government programme to fight corruption in government departments. However, corruption have experienced and practiced in the ministries that supposed to fight it.

C. Constant changing in ICT's ministers

Among the issues revealed by interviewees in which represent one of the major reasons behind the failure of the project is the continual changing of ICT ministers and officials. The interviewees explicitly stated that ministerial positions are unstable, and ministers usually stay in office for only a few months. An interviewee in the ministry of Communication and Information Technology stated:

"Up to this year (2013), I have worked with 9 ministers. Some of them give priority to developing the e-government project; while others held it back. You can notice that every minister has different priorities"

In short, ministers do not have much time to make any improvements in the project. A project of this nature (i.e. national project) needs a long time to reach completion, and consequently the constant change of officials and/or ministers in Jordan, hinders progress on the project and as a result cannot be completed.

According to [36] the continual changing in official positions over 18 months further impacted negatively to the efficiency of the e-government implementation, the case that applied to Jordan.

D. Lack of awareness on e-government

The majority of the interviewees claimed that some employees in the government department are unaware of what e-government means. One interviewee asserted:

"Some of the public sector employees believe that e-government is another name of the IT department"

Another interviewee stated:

"It is a big problem when you meet some employees who think that we are going to computerise the departments as a mean of e-government"

This result was mentioned in [37] as government officials in developing world are focusing on the technology, rather than information when implementing e-government. To conclude, there is no general agreement among employees concerning the concept e-government, and therefore government employees might deal with the project as a process of computerisation rather than dealing with it by its real nature.

V. CONCLUSION

The aim of this paper was to uncover the reasons behind what so-called total failure in e-government project in Jordan. [21, p2] categorises e-government initiatives into three camps:

- Total failure: the initiative was never implemented or was implemented but immediately abandoned.
- Partial failure: major goals for the initiative were not attained and/or there were significant undesirable outcomes.
- Success: most stakeholder groups attained their major goals and did not experience significant undesirable outcomes.

As this paper has shown that since fourteen years of launching the program, only three services have been provided electronically. Jordanian government has decided to freeze the working on e-government programme and stop all e-government projects in other ministries, until determining new to be suitable for Jordanian context [38] [39].This research has suggested, however, that government should take into consideration the above argument of why such project has failed. Without addressing the above issues, it is unclear whether Jordan will be able to implement e-government or not. Economic situation, corruption, constant changing in ICT ministers, and lake of awareness of e-government, considered by interviewees to be core factors behind the failure of e-government implementation in Jordan, as shown in this paper. Further research could be interesting to investigate how to overcome the above factors, so Jordan can be moved from total failure into partial one.

REFERENCES

[1] R. K. Kanaan, "Making sense of e-government implementation in Jordan: A qualitative investigation". Unpublished doctoral dissertation, De Montfort University, Leicester, UK, 2009.

[2] M. Yildiz, "E-government research: reviewing the literature, limitations, and ways forward". Government Information Quarterly, 24(3): pp.646-665, 2007.

[3] United Nations, "E-government survey, from e-government to connected governance". New York, 2008

[4] Z. Liao, & M.T. Cheung, "Internet-based e-banking and consumer attitudes: an empirical study". Information and Management, Vol. 39, Issue. 4, pp.283-295, 2002.

[5] C. Leitner, "E-Government in europe: the state of affairs". Presented at e-government 2003 conference, Como, Italy, 7-8 July, 2003.

[6] CapGemini and TNO Consulting, "Does egovernment pay off!" EUREXEMP- final report, final version, 2004.

[7] United Nations, "Benchmarking e-government: a global perspective", Assessing the Progress of the UN Member States. United Nations – DPEPA and ASPA, 2001.

[8] V. Ndou, "E-government for developing countries: opportunities and challenges". The Electronic Journal on Information Systems in Developing Countries. Vol.18, Issue.1, pp.1-24, 2004.

[9] [9] T. Carbo, & J. Williams, "Models and metrics for evaluating local electronic government systems and services. Electronic Journal of E-government, Vol. 2, Issue. 2 pp.95-104, 2004.

[10] M.M. Reynolds, & M. Regio, "Egovernment as a catalyst in the information age, microsoft e-government initiatives", 2001 Available at: www.netcaucus.org/books/egov2001 //pdf/EGovIntr.pdf. Date of Access 1 August 2013, 2001.

[11] M.E. Cook, M.F. LaVigne, C.M. Pagano, S.S. Dawes, & T.A. Pardo, "making a case for local e-government". Center for Technology in Government, 2002.

[12] IDABC eGovernment Observatory, "The impact of eGovernment on competitiveness, growth and jobs", 2005. Available at: http://europa.eu.int/idabc/egovo. Date of access June, 10, 2013.

[13] J. Seifert, "A primer on e-government: sectors, stages, opportunities, and challenges of online governance". Report for Congress, 2003.

[14] World Bank, "The E-government handbook for developing countries". A project of InfoDev and the Center for Democracy and Technology, 2002.

[15] K. Layne, & J. Lee, "Developing fully functional e-government: a four stage model". Government Information Quarterly. Vol. 18, Issue 2, pp.122-136, 2001.

[16] Deloitte Research, "At the dawn of e-government, the citizen as customer", 2000, Available at http://www.egov.vic.gov.au/pdfs/e-government.pdf. Date of access 1 October 2013.

[17] ESCWA, "Promoting e-government applications towards an information society in escwa member countries". Western Asia Preparatory Conference for the World Summit on the Information Society (WSIS) Beirut, 4-6 February, 2003.

[18] J.S. Hiller, & F. Belanger, "Privacy strategies for electronic government". E-Government Series. Arlington, VA: PricewaterhouseCoopers Endowment for the Business of Government, 2001.

[19] M. R. Muhammad, "Managing the implementation of e-government in malaysia: a case of e-syariah". Australian Journal of Basic and Applied Sciences, 7(8): pp.92-99, 2013.

[20] D. Dada, "The failure of e-government in developing countries: a literature review". The Electronic Journal on Information Systems in Developing Countries, 26, 7, pp.1-10, 2006

[21] R. Heeks, "Most egovernment-for-development projects fail: how can risks be reduced". iGovernment Working Paper Series. University of Manchester, 2003.

[22] CIA, "The world factbook". Available at: https://www.cia.gov/library/publications/the-world-factbook/geos/jo.html. Date of Access 1 November 2013.

[23] MOICT, "E-governnment program: vision & mission". Available at: http://www.moict.gov.jo/en-us/egovernmentprogram/visionmission.aspx Date of Access 25 November 2013.

[24] A. Al-Omari, & H. Al-Omari, "E-government readiness assessment model". Journal of Computer Science 2 (11): pp.841-845, 2006.

[25] M. Al-Shboul, & I. Alsmadi, " Jordan e-government challenges and progresses". International Journal of Advanced Corporate Learning, Volume 3, Issue 1, 2010.

[26] M. Al-Zoubi, T. Sam, & L. Eam, "E-Government adoption among businesses in Jordan". Academic Research International, Volume 1, Issue 1, 2011.

[27] M. Al-Zoubi, T. Sam, & L. Eam, "Analysis of e-government adoption and organization performance in the jordan businesses sector". Academic Research International, Volume 1, Issue 3, 2011.

[28] B.C. Stahl, "The ethical problem of framing e-government in terms of e-commerce". Electronic Journal of E-Government. Vol. 3, Issue. 2, pp.77-86, 2005.

[29] S. Khasawneh, Y. Jalghoum, O. Harfoushi, & R. Obiedat, "E-Government program in jordan: from inception to future plans". International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, 2011.

[30] F.T. Qutaishat, "Users' perceptions towards website quality and its effect on intention to use e-government services in Jordan". International Business Research; Vol. 6, No. 1, 2013.

[31] R. K. Yin, "Case study research: Design and methods". California, USA, Sage Publications. Thousand Oaks, 2003.

[32] A. Strauss, & J. Corbin, "Basics of qualitative research: grounded theory procedures and techniques", Sage, London, 1990.

[33] C. Fidler, R. Kanaan, & S. Rogerson, "Barriers to e-government implementation in jordan: the role of wasta". International Journal of Technology and Human Interaction, Vol. 7, No. 2 April-June, 2011.

[34] J. Choudrie, V. Weerakkody & S. Jones, "Realizing e-government in the uk: rural and urban challenges". The Journal of Enterprise Information Management, Vol. 18, Issue. 5, pp.568-585, 2005.

[35] Transparency International, "Transparency international; the global coalition against corruption". Available at: http://www.transparency.org/country#JOR_DataResearch_SurveysIndices. Date of Access 25 November 2013.

[36] J. Seifert, & G. Bonham, "The transformative potential of e-government in transactional democracies". Washington, D.C., Congressional Research Service, 2004.

[37] United Nation, "UN global e-government readiness report, toward access for opportunity". New York, 2004.

[38] Sarayanews, "Online newspaper". Available at http://www.sarayanews.com/index.php?page=article&id=215947. Date of Access 21 November 2013.

[39] Maqar News, "Online newspaper" Available at http://maqar.com/?id=31794 Date of Access 21 November 2013.

Achieving Regulatory Compliance for Data Protection in the Cloud

Enterprise Approaches to Distributed Encryption Management

Mark Rivis, Shao Ying Zhu
School of Computing and Mathematics,
The University of Derby,
Derby, United Kingdom.

***Abstract*—The advent of cloud computing has enabled organizations to take advantage of cost-effective, scalable and reliable computing platforms. However, entrusting data hosting to third parties has inherent risks. Where the data in question can be used to identify living individuals in the UK, the Data Protection Act 1998 (DPA) must be adhered to. In this case, adequate security controls must be in place to ensure privacy of the data. Transgressions may be met with severe penalties. This paper outlines the data controller's obligations under the DPA and, with respect to cloud computing, presents solutions for possible encryption schemes. Using traditional encryption can lead to key management challenges and limit the type of processing which the cloud service can fulfill. Improving on this, the evolving area of homomorphic encryption is presented which promises to enable useful processing of data whilst it is encrypted. Current approaches in this field have limited scope and an impractical processing overhead. We conclude that organizations must thoroughly evaluate and manage the risks associated with processing personal data in the cloud.**

Keywords—cloud computing; data protection legislation; Data Protection Act 1998; homomorphic encryption; data privacy; symmetric encryption

I. INTRODUCTION

In the UK, information relating to living individuals is regulated by some of the most rigorous privacy legislation in the world. The Information Commissioner's Office (ICO) regularly has cause to use its powers to enforce the Data Protection Act 1998 (DPA). Punitive fines can be imposed, often running to hundreds of thousands of pounds, depending on the nature and impact of the data breach [8]. There is at present no duty to disclose a data breach to the ICO [9]; however the obligation to adhere to the DPA is a serious one, with severe penalties for non-compliance [10].

Whilst managing their statutory obligations, most organizations are also duty bound to make cost effective use of their computing resources. One means of reducing costs is the move from traditionally hosted computing services to those present "in the cloud". Cloud computing can offer flexibility, convenience and resilience for essential business services, usually at a reduced total cost [2].

Martens, Walterbusch and Teuteberg [22] present a model to assess the total cost of ownership (TCO) of cloud computing services. By taking advantage of shared environments in a multitenant model, computing service providers can deliver systems over the Internet which often makes a compelling proposition.

However, as an evolving computing paradigm, the risks to data privacy must be carefully balanced with any *prima facie* benefits of cloud computing. Initial deployments of cloud computing systems have been secured using existing technology, such as secure sockets layer (SSL) and public key encryption. However, as the technical vulnerabilities of a shared data processing infrastructure become better known, it is apparent that new means of securing data in this environment must be sought.

Whether any business will take up public cloud computing services depends primarily on their appetite for risk, versus the expected cost savings. In many cases businesses are avoiding the uncertainty of using cloud services in preference for a more assured security posture [1]. To fully enable the take up of cloud computing, new approaches to storing and processing encrypted data in a shared environment must be developed.

This paper's contribution is in presenting details of UK privacy legislation in the context of cloud computing. Detailed research has been carried out in the computer science literature on the latest techniques for processing encrypted data in a shared computing environment. A critique of these approaches is offered and the conclusion drawn on the efficacy of these approaches. Final comments are made on the open issues which will be the subject of future work.

The rest of this paper is structured as follows. The background to current data protection legislation is presented in Section 2, with specific reference to the security principles of the DPA reviewed in Section 3. Concepts of cloud computing are discussed in Section 4. Following, in Section 5, the applicability of the DPA and implications of a data breach are considered. Measures to achieve data compliance are presented in Section 6 with a review of approaches to encryption. Anticipated future developments are discussed in Section 7, and final conclusions are drawn in Section 8.

II. DATA PROTECTION LEGISLATION

Legislative instruments are in place in parts of the western world giving individuals certain rights over the data which is held about them. Specifically in the European Union (EU), the Data Protection Directive of 1995 [23] required member states to enact their own local legislation, giving a consistent

approach to data protection across nations. As a consequence of this, the UK enacted the DPA [5]. It is this act which provides the basis for a discussion of data protection here, but of course the principles could apply in parallel across other EU member states.

The United States has no single piece of legislation providing a comparative basis for data protection. Instead, a "sectoral" approach is taken, with different business areas required to comply within their own framework of regulation. Examples are the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [25], and Federal Information Security Management Act of 2002 (FISMA) [26], which apply to the health insurance industry and federal agencies respectively.

Given the differing approaches to privacy legislation, there would not ordinarily be a simple method of comparing the levels of data protection afforded by various industries in the US and EU. To smooth the path to global trade (where the sharing or export of data is required), the US and EU have worked together to develop the "Safe Harbor" scheme. The EU regulations have stricter privacy controls in general, so US companies can be assessed for Safe Harbor, which ensures they meet the EU "adequacy" requirement for data privacy [21].

III. United Kingdom Data Protection Act 1998

A brief summary of the DPA is presented here, with particular note being made of the act's relevance to cloud computing. All references are drawn from the online record of the legislation published by Her Majesty's Stationery Office [5].

The scope of the DPA is restricted to systems processing personal information which could be used to identify a living individual. With respect to cloud computing (or any other computing platform), this initial test of applicability will determine whether regulatory compliance is required for that system.

Three roles are defined by the act. They are the data subject (the person to whom the data relates), the data controller (who determines the purpose and manner of data processing) and the data processor (who carries out processing of the data on behalf of the data controller). In a cloud computing environment it is most commonly the case that the data processor will be a third-party company providing cloud services. It is essential then that the data controller and data processor have a clear understanding of the nature of their relationship, and their obligations under the act.

The act lists eight principles which define how data may be processed [24]. These can be summarized as follows:

1) Personal data shall be processed fairly and lawfully

2) Personal data shall be obtained and processed only for one or more specified and lawful purposes

3) Personal data shall be adequate, relevant and not excessive

4) Personal data shall be accurate and kept up to date.

5) Personal data shall be kept only for as long as necessary to fulfil the purpose for which it was obtained

6) Personal data shall be processed in accordance with the rights of data subjects under this Act

7) Appropriate security measures must be taken to prevent unauthorized access to data, and to prevent accidental loss, destruction or amendment of personal data.

8) Personal data shall not be transferred to any country outside the European Economic Area (EEA), unless that country has adequate levels of protection for processing personal data.

Most of the principles should be met through the operational business processes of the data controller. It can be seen that purely technical measures will not enable compliance with the act in respect of many of these principles. However it is principles seven and eight which warrant closer scrutiny in a cloud computing environment.

IV. Cloud Computing

A. Definition

Often described as an emerging computing paradigm, cloud computing is enabled by the conjunction of multiple maturing technologies. For some years it has evaded a clear and concise description; however this has evolved as the technologies used have become more established. A definition of cloud computing is provided by the National Institute for Standards and Technology (NIST) in their Special Publication 800-145 [14]. Five key attributes are present in their definition:

- On-demand self-service – the consumer is able to provision the service autonomously
- Broad network access – access is available from heterogeneous devices and network media
- Resource pooling – a multi-tenant model enables economies of scale for the service; location independence is often a feature here but in some cases the customer can restrict the physical location of their resources (e.g. by country or data centre)
- Rapid elasticity – depending on demand the resource available can be scaled up or retracted
- Measured service – the model enables pay-per-use accounting, with only the resources used being charged for.

As well as the defining cloud computing characteristics, [14] also offers definitions of three service models:

Software as a Service (SaaS) – access to an application is provided as part of the service. The customer has no control over the applications infrastructure or how it operates, other than processing their data (which may entail choosing some application level settings.) An example is the Customer Relationship Management (CRM) system provided by salesforce.com.

Platform as a Service (PaaS) – the consumer is able to deploy applications to the cloud infrastructure but is not able to influence the underlying architecture. System components can be implemented as middleware, harnessing the cloud-service processing power as they interact with other components. An

example is Google's AppEngine service where a range of web applications can be integrated using Application Programming Interfaces (APIs.)

Infrastructure as a Service (IaaS) – in this model the capability to provision lower level system components is provided to the user. Decisions can be made as to the size and type of resources such as processors, memory, network and storage. They may usually choose the type of operating system and application software which is deployed to the infrastructure. Amazon's Elastic Compute Cloud (EC2) is a prime example of an IaaS offering. Weinhardt, et al., [29] provided an early representation of this model in their Cloud Business Model Framework. Their work highlights the evolution of cloud computing, differentiating it from the earlier grid computing paradigm. A graphical representation of this is reproduced in Figure 1, which summarizes this layered approach to cloud platforms [29].

B. Data Protection Implications in Cloud Computing

Drawing together the requirements of the DPA and the definition of cloud computing, it can be seen that some challenges and risks to compliance in this environment are brought about. Introducing a third party as a cloud service provider changes the data processor role as defined by the act. Therefore the relationship between the data controller and data processor must be clearly defined and respective responsibilities well understood. A written contract is required by the DPA between the data controller and the data processor, requiring that "the data processor is to act only on instructions from the data controller" and "the data processor will comply with security obligations equivalent to those imposed on the data controller itself" [7]. In this guidance, the ICO also describes how organizations should be aware of the possible "layered" nature of cloud computing services. That is to say, a data controller may use an on-line web application from cloud provider A (as SaaS), however that provider may also be using cloud provider B to host its services (as IaaS.)

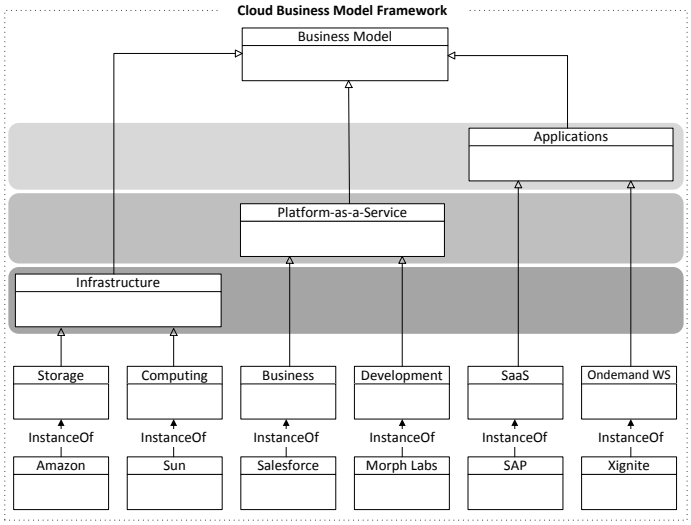


Fig. 1. Cloud Business Model Framework

These potentially complex relationships and roles (with respect to the DPA) must be well understood.

Principle eight of the DPA says that personal data may not be exported outside of the EEA, unless adequate data protection controls are in place. As noted, US companies may be accredited for the Safe Harbor scheme to demonstrate that their controls are adequate. Therefore when choosing a cloud service provider it is essential to understand broadly where personal data will be stored, restricted to territories within the EEA or others with "adequate" data privacy controls. This must be the subject of binding service agreement such that the data controller's obligations under the DPA can be met.

V. Implications of a Data Breach

Aside from the short term effect on staff productivity, the primary implications of a data breach against a business or other organization are twofold.

Firstly, if the entity was deemed to be negligent in contravening the DPA, the ICO has the power to act in some way. Its range of powers includes the criminal prosecution of those who contravene the act, and the issuing of monetary penalties of up to £500,000 for serious breaches. This would be the case where the security measures in place were not sufficient, counter to principle 7 of the act. It is recognized however that not all data breaches are equal; the ICO will take a different view of a breach due to poor procedural controls on the part of the data controller than it would where a criminal act is committed by a malicious attacker. Guidance published by law firm Pinsent-Masons states "a data controller will not be in breach of the Act if it can show that appropriate security measures were taken" [15].

The second aspect of a data breach is the reputational damage which follows if (or when) the event becomes public knowledge. Note that there is currently no requirement in the UK for organizations to report data breaches [9]. Future legislation may change this. Draft EU legislation is being discussed which would enforce the requirement to notify local data protection authorities and the data subjects of a breach [3]. The reputational impact may have varied effects on the organization, including reduced customer loyalty, reduced amount of new business, increased customer turnover and potentially a reduced share price for stock market listed companies [16].

Clearly the impact of a publicly known data breach goes beyond the lost effort of resolving the immediate issues, and the financial penalty imposed by regulatory bodies. Organizations must therefore balance the risk of a data breach occurring, with the cost of any measures which are put in place to prevent such breaches.

VI. Measures to Achieve DPA Compliance in Cloud Computing

A. Symmetric Encryption

The use of encryption is recognized as a means to achieving DPA compliance by ensuring that only authorized parties can read the personal data in question. The suitability of solutions for encrypting data in the cloud depends on various factors

which include, *inter alia*, the number of users requiring access, the nature of data sharing between users, the level of computation the cloud system is required to perform and the quantity and structure of data in question.

In proposing a simple encryption scheme to protect data, a further challenge arises, since encrypted data cannot normally be processed in the cloud (as though it were stored in the clear.) Whilst this approach provides a good degree of security assurance, any agent that does not have the necessary key to decrypt the data is unable to interact with it. In this sense, the cloud is reduced to providing data storage at a third-party facility. This may be desirable in some cases, but the capability of providing some kind of remote processing of data may be required.

Puttaswamy, Kruegel and Zhao [17] identify that when processing data it is often the case that the actual value of some data elements does not need to be known. It is therefore feasible to encrypt some data and still carry out useful operations on it. This type of data they describe as functionally encryptable, that is, data which can be encrypted without affecting the functionality of the application. In their system "Silverline", they describe a multistage framework to achieve this. The method includes automatically identifying functionally encryptable data, managing keys to encrypt and decrypt data shared within role-based groups of users, and providing transparent data access through a scheme of trusted and untrusted web browser interface components.

Their results show that for some common web applications about 70-80% of data could be encrypted by this system. The focus on automation is designed to allow an easy transition for existing web applications to using this encryption model in the cloud.

There is however a significant drawback in the approach of partial encryption. Whilst it is convenient to take an existing application and automatically modify its behaviour to encrypt some of its data operations, there is no assurance that personal data would be considered functionally encryptable for the application in question. In that case there is no advantage to using the Silverline system as a means of increasing security to achieve DPA compliance.

A system relying on symmetric encryption key management is proposed by Litwin, Jajodia and Schwarz [11]. In their scheme, each user has their own symmetric key which is used to encrypt data uploaded to the cloud. All symmetric keys are also uploaded to the cloud but remain hidden by some means. If a user wants to share data with another, they unhide their symmetric key so that they are both able to decrypt the data. This approach, whilst sound in so far as can be applied, is limited in its potential. It works along the lines of traditional file sharing permissions, applying encryption, and using key management as a method of granting and revoking access. It does not leverage the computing capabilities of cloud systems beyond data storage and sharing.

An alternative approach to enabling cloud computing for security-sensitive businesses relies on separating data storage and data processing providers [28]. Interfacing between these services is an independent encryption/decryption service in the cloud. This approach enables service providers to be selected based on differing levels of trust, with processing being handled at a highly trusted cloud service, and (encrypted) data storage at a less trusted facility.

B. Homomorphic encryption

It is recognized in the field of cryptanalysis that the ability to process data whilst it is encrypted would open the way to new distributed computing models, with reduced risks to data privacy. In effect, the data processor would be able to carry out some operations on the data without actually being able to decrypt it. This was recognized by Rivest, Adleman and Dertouzos in [18] where a set such functions is described. Their work is an extension of the earlier paper which described the RSA public key cryptosystem [19]. The correlation comes about because RSA encryption is homomorphic for multiplication (but not addition).

Homomorphic operations can be carried out on data without decrypting it in the conventional sense. An inherent limitation is noted in [18], however, in that comparison operations would not be allowed under such a model. If that were the case, then a malicious user would be able to carry out a simple binary search of encrypted data, and once a comparison match was found, would be able to deduce the unencrypted value. In identifying the possibility of privacy homomorphisms, Rivest, Adleman and Dertouzo [18] opened the way for further work to find computationally practical implementations.

A fully homomorphic encryption (FHE) scheme is one where arbitrarily complex operations can be carried out against encrypted data.

Gentry [4] described such a scheme based on ideal lattice cryptography. It is recognized however that this is computationally impractical as the amount of effort required increases rapidly as the security level increases.

Attempts to refine and enhance FHE schemes continue. In addition to new cryptanalysis techniques, some trials attempt to speed up the encryption and decryption processes with massively parallel programming techniques using graphics processing units (GPUs). Single-digit improvements in performance are reported [27]. Whilst providing some improvement, this does not in itself make such techniques practically useful.

A somewhat homomorphic encryption (SwHE) scheme can carry out some set of defined operations on encrypted data. Naehrig, Lauter and Vaikuntanathan [12] describe a SwHE scheme based on the "ring learning with errors" problem, which exhibits good performance. They also describe some real-world cloud computing scenarios where the limited functionality of SwHE could be employed.

Research continues into developing a comprehensive fully homomorphic encryption system which can be implemented using practical computation resources. López-Alt, Tromer and Vaikuntanathan [13] describe how multiple parties can encrypt their own data for collaborative processing in the cloud, whilst retaining privacy over their original data. This is implemented using an enhancement of NTRU encryption, an efficient public

key cryptosystem based on lattices, originally developed by Hoffstein, Pipher and Silverman [6]. NTRU is fast and scalable when compared to other public key crypto systems; it is claimed that for increases in the message length *n*, the encryption and decryption requirements of NTRU increases with n^2, where RSA increases with n^3 [6].

When encryption is applied to cloud computing systems the matter of key management becomes problematic. Some challenges around user access management become apparent in this case:

- How do we revoke access to data for some users without having to re-encrypt data?
- How do we avoid collusion between users and the cloud provider?
- How are changes to user privileges managed?

Samanthula, Howser, Elmehdwi and Madria [20] propose a scheme to handle such key management challenges using homomorphic encryption and proxy re-encryption, whereby encrypted data is re-encrypted for a new recipient without having to decrypt it first. This approach gives some benefits in handling user management for data sharing but their focus neglects the need to process data in the cloud. For example, if encrypted data is written to a database, the database process itself would either need to decrypt the data or carry out FHE operations on it to do any useful work. In the former case the same risk is present, in that the database process may be compromised allowing a malicious attacker to read the data. It is the latter case, that is, finding a workable FHE scheme which remains elusive.

The future implications of being able carry out computational operations on encrypted data are not yet clear. It seems likely that if an encryption scheme is available which has low management overhead, yet retains privacy amongst untrusted third party service providers, the potential for enterprise take up of cloud computing will greatly increase. However the current restriction of being unable to carry out comparison functions seems to limit the actual usefulness of this approach in processing information.

VII. Future Work

In the field of cryptography, work continues with the focus of finding either a FHE scheme which is computationally efficient, or a SwHE scheme which is sufficient to meet specific use cases. In both scenarios the goal is to find a system which is effective and has a manageable overhead in terms of its operational processes, for example, key management.

When applied to cloud computing, techniques which rely on encrypting partial data attributes or managing symmetric encryption keys require a certain amount of application customization. They may provide some assurance that DPA compliance can be achieved but the development overhead means such approaches are largely bespoke.

For organizations concerned with DPA compliance, current approaches to cloud computing are primarily based on risk management. This requires consideration of factors in three areas.

Technological risks of using cloud systems. Resources are available which highlight some technological risks [1] but this is an area which continues to evolve. Periodic assessment of the weaknesses of cloud computing and the capabilities of threat agents should be undertaken.

Legal obligations. Notwithstanding the operational and reputational impacts of a data breach, there is a requirement to provide security which is 'good enough' for the organization. Measures must be taken to ensure data privacy based on the perceived risks and appropriate to the nature of the data.

Organizational relationships. Cloud computing simplifies the deployment of technology, but adds complexity in organizational relationships. Multiple third parties may be involved, either at inception or in the future. There must be a clear understanding of the responsibilities and obligations of all parties.

VIII. Conclusion

Cloud computing presents a valuable opportunity to organizations, enabling services which are not only cost effective, but flexible, resilient and accessible from anywhere on the Internet.

There is of course no variation in the legal framework with respect to cloud computing; organizational obligations remain consistent in this regard. Clarification on the applicability of the DPA is available from the ICO and supporting resources are published by proponents of cloud initiatives.

Encryption of sensitive data is recommended by the ICO as a means to preserve data privacy. However using traditional methods limits the applications where cloud computing can be used effectively, and presents major challenges in key management.

The evolving field of homomorphic encryption appears to offer promising opportunities here. By using specific encryption schemes, it is possible to carry out useful operations on encrypted data without ever having access to the unencrypted data. Early work has demonstrated the principle, but resulted in impractical computational workloads. The secure processing of encrypted data by a third party is not yet practicable.

Organizations must be cognisant of the implications of using cloud computing services, taking a risk based approach specific to their circumstances.

References

[1] Cloud Security Alliance, 2010. Top Threats to Cloud Computing V1.0. Available at: https://cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf, accessed 2 December 2012.

[2] Cloud Security Alliance, 2011. Security Guidance for Critical Areas of Focus in Cloud Computing v3.0. Available at: https://cloudsecurityalliance.org/guidance/csaguide.v3.0.pdf, accessed 2 December 2012.

[3] European Parliament, 2009. Directive 2009/136/EC of the European Parliament. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:337:0011:01:EN:HTML, accessed 6 December 2012.

[4] C. Gentry, "Fully homomorphic encryption using ideal lattices," Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC '09), ACM, pp. 169-178, Jun. 2009, doi:10.1145/1536414.1536440.

[5] Her Majesty's Stationery Office, n.d. Data Protection Act 1998. Available at: http://www.legislation.gov.uk/ukpga/1998/29/contents, accessed 2 December 2012.

[6] J. Hoffstein, J. Pipher, J. H. and Silverman, "NTRU: A ring-based public key cryptosystem," Proceedings of the Third International Symposium on Algorithmic Number Theory (ANTS-III), Springer-Verlag, Jun. 1998, pp. 267-288, doi:10.1007/BFb0054868.

[7] Information Commissioner's Office, 2012. Guidance on the Use of Cloud Computing. Available at: http://www.ico.gov.uk/for_organisations/guidance_index/~/media/documents/library/Data_Protection/Practical_application/cloud_computing_guidance_for_organisations.ashx, accessed 7 December 2012.

[8] Information Commissioner's Office, 2012. Monetary penalty notices. Available at: http://www.ico.gov.uk/enforcement/fines.aspx, accessed 7 December 2012.

[9] Information Commissioner's Office, 2012. Notification of data security breaches to the Information Commissioner's Office (ICO). Available at: http://www.ico.gov.uk/for_organisations/guidance_index/~/media/documents/library/Data_Protection/Practical_application/breach_reporting.ashx, accessed 6 December 2012.

[10] Information Commissioner's Office, n.d. Taking action: data protection and privacy and electronic communications. Available at: http://www.ico.gov.uk/what_we_cover/taking_action/dp_pecr.aspx, accessed 6 December 2012.

[11] W. Litwin, S. Jajodia, and T. Schwarz, "Privacy of data outsourced to a cloud for selected readers through client-side encryption," Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society (WPES '11), ACM, Oct. 2011, pp. 171-176, doi:10.1145/2046556.2046580.

[12] M. Naehrig, K. Lauter, and V. Vaikuntanathan, "Can homomorphic encryption be practical?," Proceedings of the 3rd ACM Workshop on Cloud Computing Security Workshop (CCSW '11), ACM, Oct. 2011, pp. 113-124, doi:10.1145/2046660.2046682.

[13] A. López-Alt, E. Tromer, and V. Vaikuntanathan, "On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption," Proceedings of the 44th Symposium on Theory of Computing (STOC '12), ACM, May 2012, pp. 1219-1234, doi:10.1145/2213977.2214086.

[14] National Institute of Standards and Technology, 2011. The NIST Definition of Cloud Computing. Available at: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf, accessed 2 December 2012.

[15] Pinsent Masons, n.d. Data Security. Available at: http://www.pinsentmasons.com/en/media/published-articles/data-security/, accessed 6 December 2012.

[16] Ponemon Institute LLC, 2012. Aftermath of a Data Breach Study. Available at: http://www.experian.com/assets/data-breach/brochures/ponemon-aftermath-study.pdf, accessed 6 December 2012.

[17] K. P. Puttaswamy, C. Kruegel, and B. Y. Zhao, "Silverline: toward data confidentiality in storage-intensive cloud applications," Proceedings of the 2nd ACM Symposium on Cloud Computing (SOCC '11), ACM, Oct. 2011, article no. 10, doi:10.1145/2038916.2038926.

[18] R. L. Rivest, L. Adleman and M. L. Dertouzos, "On data banks and privacy homomorphisms," in "Foundations of Secure Computing," pp. 169-180, edited by DeMillo, R., Dobkin, D., Jones, A. and Lipton, R., New York: Academic Press, 1978.

[19] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," Communications of the ACM, vol. 21, iss. 2, Feb. 1978, pp. 120-126, doi:10.1145/359340.359342.

[20] B. K. Samanthula, G. Howser, Y. Elmehdwi, and S. Madria, "An efficient and secure data sharing framework using homomorphic encryption in the cloud," Proceedings of the 1st International Workshop on Cloud Intelligence (Cloud-I '12), ACM, Aug. 2012, article no. 8, doi:10.1145/2347673.2347681.

[21] US Department of Commerce International Trade Administration, 2012. Welcome to the U.S.-EU Safe Harbor. Available at: http://export.gov/safeharbor/eu/eg_main_018365.asp, accessed 7 December 2012.

[22] B. Martens, M. Walterbusch, F. Teuteberg, "Costing of cloud computing services: a total cost of ownership approach," Proceedings of the 2012 45th Hawaii International Conference on System Sciences (HICSS '12), IEEE Computer Society, Jan. 2012, pp. 1563-1572, doi:10.1109/HICSS.2012.186.

[23] European Parliament, 1995. Directive 95/46/EC of the European Parliament. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML, accessed 4 February 2013.

[24] Information Commissioner's Office, n.d., Data protection principles. Available at: http://www.ico.gov.uk/for_organisations/data_protection/the_guide/the_principles.aspx, accessed 4 February 2013.

[25] US Government Prinitng Office, n.d., Public Law 104 - 191 - Health Insurance Portability and Accountability Act of 1996. Available at: http://www.gpo.gov/fdsys/pkg/PLAW-104publ191/content-detail.html, accessed 4 February 2013.

[26] The Library of Congress, n.d., E-Government Act of 2002. Available at: http://thomas.loc.gov/cgi-bin/query/z?c107:H.R.2458.ENR:, accessed 4 February 2013.

[27] Wei Wang; Yin Hu; Lianmu Chen; Xinming Huang; B. Sunar, "Accelerating fully homomorphic encryption using GPU," High Performance Extreme Computing (HPEC), 2012 IEEE Conference on, 10-12 Sept. 2012, pp.1-5, doi: 10.1109/HPEC.2012.6408660

[28] Jing-Jang Hwang; Hung-Kai Chuang; Yi-Chang Hsu; Chien-Hsing Wu, "A Business Model for Cloud Computing Based on a Separate Encryption and Decryption Service," Information Science and Applications (ICISA), 2011 International Conference on, 26-29 April 2011, pp.1-7, doi: 10.1109/ICISA.2011.5772349

[29] C. Weinhardt, A. Anandasivam, B. Blau, N. Borissov, T. Meinl, W. Michalk, and J. Stößer,"Cloud computing – a classification, business models, and research directions," Business and Information Systems Engineering (BISE), vol. 1, no. 5, pp. 391-399, 2009.

Pre-Eminance of Open Source Eda Tools and Its Types in The Arena of Commercial Electronics

Geeta Yadav
VLSI Design Group
Department of EECE, ITM University
Gurgaon (Haryana), India

Neeraj Kr. Shukla
VLSI Design Group
Department of EECE, ITM University
Gurgaon (Haryana), India

***Abstract*—Digital synthesis with a goal of chip designing in the commercial electronics arena is packed into large EDA Software providers like, Synopsys, Cadence, or MentorGraphics. These commercial tools being expensive and having closed file structures. It is also a financial constraint for the startup companies sometimes who have their budget limitations. Any bug-fixes or add features cannot be made with ease; in such scenario the company is forced to opt for an alternative cost effective EDA software. This paper deals with the advantages of using open source EDA tools like Icarus Verilog, Verilator, GTKwave viewer, GHDL VHDL simulator, gEDA, etc. that are available as a free source and focuses on the Icarus Verilog simulator tool. It can be seen as a big encouragement for startups in Semiconductor domain. Thereby, these open source EDA tools make the design process more cost-effective, less time consuming and affordable as well.**

***Keywords*—Open source EDA; MentorGraphics; Cadence; Icarus Verilog; GTKwave viewer; Verilator; GHDL VHDL simulator; gEDA; Linux; Github; VPI**

I. INTRODUCTION

A. Open source EDA tools

The open source plays a key role in the EDA tool development. A set of tools known as Digital synthesis flow is used to turn a circuit design written in Verilog or VHDL a high-level behavioral language into a physical circuit, which can either serve to be configuration code for a Xilinx or Altera chip, or a layout in a specific fabrication process technology [1]. In the commercial electronics arena, digital synthesis with the application of a chip design is usually packaged into large EDA software systems like MentorGraphics or Cadence or Synopsys which are very expensive. So the designers need to maintain cutting-edge performance as these commercial tool chains get more and more expensive. Another disadvantage of working with the closed file structures is that it becomes difficult to add customized features to the tool. Also, for small customers it becomes very difficult and time consuming for them to fix the bugs appearing in the tool. An alternate to this, is to go for new software that is more user friendly and easier to deal with [2].

The oldest of these are probably VIS and SIS, two software tools developed at Berkeley for Verilog parsing, logic verification, and mapping of logic onto a digital standard cell library. They perform the task of logic optimization and cell mapping admirably [3].

B. Linux Platform

Linux serves as a open source platform for the various EDA tools. It is very advantageous as compared to the other operating systems as it is free to obtain, while Microsoft products are available for a hefty and sometimes recurring fee, the security aspect of Linux is much stronger than that of Windows as free from virus, the power to control just about every aspect of the operating system, flexibility, its use as a firewall, a file server, or a backup server [4].

C. Github

It is a web based hosting service which is used for software development projects using the Git version control system. Git is a software version control tool which is mainly used for proper management of the various versions of a particular tool. It keeps track of the dates when the changes are being made to the tool, what are those changes, and who has made those changes. This way, we can easily follow the continuous changes being made to the tool. For private repositories, Github provides paid plan whereas for open source projects it is free. The site provides social networking functionality such as feeds, followers and the social network graph to display how developers work on their versions of a repository.

II. TYPES OF OPEN SOURCE EDA TOOLS

A. Icarus Verilog Simulator

Icarus verilog is an open source synthesis and simulation tool. It works as a compiler and compiles the source code written in Verilog(IEEE-1364) into a target format. The Icarus Verilog compiler is written by Stephen Williams. He is still working on it. This tool supports a waveform viewer named GTKWave. This tool has various released versions; one of its latest released versions is version 0.9.6 [5].

Characteristics:

- Simulation engine is efficient
- Portable compiler
- Challenge for commercial tools
- Supported graphics tool like GTKwave
- New compatibility with de facto standards such as library formats and command files

Significance:

The addition of the functionality to the Icarus Verilog tool would lead to the effective and reduction in the design and verification of the circuits through the Hardware Description Languages. List of providers who offer commercial support for Icarus Verilog and/or related products are Dolly Software Private Limited, Embecosm, OCLogic Limited [6].

The tool has various release versions like 0.9.2, 0.9.3, 0.9.4, and so on, latest being the version 0.9.6. The release notes of the tool lists the various bugs in that version which are then worked on. This can be fixed by any user as it is an open source EDA detail. The fixation of these bugs leads to making the verification of the tool more effective.

B. Verilator

Verilator is a free Verilog HDL simulator. It compiles synthesizable Verilog into an executable format and wraps it into a SystemC model. Internally a two-stage model is used. The resulting model executes about 10 times faster than standalone SystemC. Verilator has been used to simulate many very large multi-million gate designs with thousands of modules. Therefore we have chosen this tool to be used in the verification environment for the Open RISC processor [7].
Characteristics:

- Verilator is the fastest
- It compiles synthesizable Verilog (not test-bench code), plus some PSL, SystemVerilog and Synthesis assertions into C++ or SystemC code
- Verilator has been used to simulate many very large multi-million gate designs with thousands of modules

C. GHDL VHDL simulator

GHDL implements the VHDL87 (common name for IEEE 1076-1987) standard, the VHDL93 standard (aka IEEE 1076-1993) and the protected types of VHDL00 (aka IEEE 1076a or IEEE 1076-2000). The VHDL version can be selected with a command line option [7].

Characteristics:

- It has been successfully employed for compiling and simulating the DLX processor and the LEON1 SPARC processor
- It directly creates binaries or executable images, which is the best form for testbenches
- It can be used to pretty print or to generate cross references in HTML

D. gEDA

The gEDA project was started by Ales Hvezda in an effort to remedy the lack of free software EDA tools for Linux/Unix. The first software was released on 1 April 1998, and included a schematic capture program and a netlister. At that time, the gEDA project website and mailing lists were also set up.

Characteristics:

- Originally, the project planned to also write a PCB layout program
- The ability to target netlists to PCB was quickly built into the gEDA Project's netlister, and plans to write a new layout program from scratch were scrapped
- The authors of other open source programs became affiliated with the gEDA website and mailing lists, and the collaborative gEDA Project was born

E. GTKwave viewer

It is a GTK+ based wave viewer for Unix, Win32 etc. GTK+ is basically a toolkit used for creating graphical user interfaces. It helps in viewing the VCD files. The Icarus Verilog tool uses the GTKwave tool for the graphical representation of the results.

III. Verilog Procedural Interface

VPI stands for Verilog Procedural Interface. The use of VPI permits behavioral Verilog code to invoke C functions and C functions to invoke standard Verilog system tasks. VPI is a part of the programming language interface standard IEEE 1364. Users can access information contained in a Verilog design as well as provides facilities to interact dynamically with a software product via the VPI interface routines. VPI interface provides applications such as annotators and connection of a Verilog simulator with other simulation tools and customize the debugging of tasks, delay calculators. The basic functions of the VPI interface can be categorized into two main areas: VPI callbacks usage for dynamic software product interaction. Verilog HDL objects and simulation specific objects can be accessed.

Callbacks in VPI can call a user-defined application by the use of Verilog HDL software product like logic simulator when a specified activity occurs on request of the user. Dynamic software product interaction is accomplished by registered callback mechanism. To understand this take an example like, when a particular net value changes the user can request the calling of the user application my_monitor and can call my_cleanup() routine when the execution of the product of the software completes. The detection of the changes in the occurrence of values, termination of simulation, time advancement, etc. can be detected by using VPI callback facilities which allows the dynamic interaction with software product. The callback feature allows applications of integration with other simulation systems, specializing of the timing checks, complexity of debugging features, etc. There are four basic reasons for callbacks as listed below [9]:

- Simulation event (e.g., change in the value on a net or a behavioral statement execution)
- Simulation time (e.g., the termination of a time queue or after certain amount of time)
- Simulator action/feature (e.g., the end of compilation, end of simulation, restart, or enter interactive mode)
- User-defined system task or function execution

Steps required writing a C function and interfacing it with a Verilog simulator:

- Write a function in C

- Associate the C function with a new system task
- Register a new system task
- Invoke system tasks

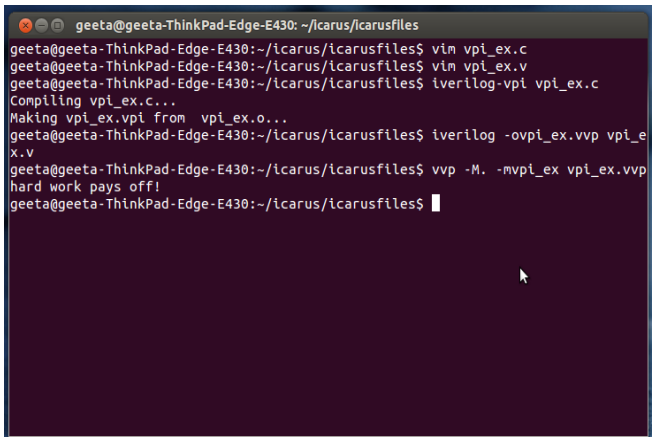


Fig. 1. Example showing use of VPI

IV. Event Queue Standard

The Verilog HDL is defined in terms of a discrete event execution model. A design has a large number of threads connections for execution or processes. Objects that can have state can be evaluated and respond to the changes in the outputs with respect to the changes in the inputs are called Processes. Modules, primitives, initial and always procedural blocks, continuous assignments, asynchronous tasks, and procedural assignment statements are all included in a process.

Update event: Change in value of a every net or variable in the circuit being simulated and named event as well.

Evaluation event: Processes have sensitivity for update events. All the processes that are sensitive to that event are evaluated in an arbitrary order when an update event is executed. Process evaluation is also an event.

Simulation time: It is the time value maintained by the simulator for modeling the actual time it would take for the circuit in simulation.

Scheduling an event: Events occur at different times slots, so in order to keep track of the events and to have a surety that they are processed in the correct order, the events are kept on an event queue, as ordered by simulation time. Putting an event on the queue is called scheduling an event.

V. VPI For Inertial Delay In Icarus Verilog Version 0.9.6

This is one of the functional bugs in the 0.9.6 released version. VPI (Verilog Procedural Interface) is used to invoke the C functions from Verilog or vice-versa. Inertial delay is basically the gate delay associated with the design. Icarus Version 0.9.6 does not provide the VPI support for inertial delays i.e. it does not consider the inertial delays associated with the design when interfacing with the other tools.

For the addition of this functionality changes need to be made in the source code. For this bug the scheduler.cc is the part of the source code where changes are needed. An event queue is used for scheduling the active, inactive, and non-blocking and monitors assignments. Hence for inertial delay a second event queue needs to be added providing input to the event loop in the scheduler.cc which will automatically keep a check on the inertial delay of the design. The adding of this functionality leads to the more efficient design verification as the inertial delay will also be encountered.

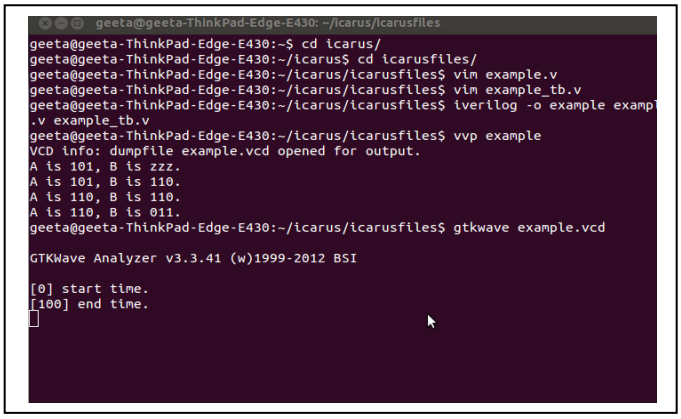


Fig. 2. Example showing inertial delay support without VPI

The inertial delay support is present in the Icarus Verilog version 0.9.6 without the use of VPI, whereas when the VPI is used than the inertial delay is not supported by the tool. Hence this is one of the bug present in the version 0.9.6.

VI. Conclusion

The closed file structures are a real challenge for the startup companies as they are expensive and the bug-fixes as well as the adding of functionality is not possible, in such scenario the company needs to change the whole software. On the other hand open source EDA tools are of great importance, it provides its source code for the users to make changes and use it.

The availability of the source code leads to the faster development of the tool. Icarus Verilog is a very strong simulation tool; the synthesis part is being worked on. Icarus Veilog also has a graphical support in the form of GTKwave viewer. So, open source EDA tools are cost-effective, less time consuming, user friendly with a lot of fun learning as well.

References

[1] URL: http://opencircuitdesign.com/qflow/welcome.html
[2] URL: http://blog.engineersimplicity.com/2008/11/open-source-eda.html
[3] URL: http://opencircuitdesign.com/qflow/welcome.html
[4] URL:http://ubuntu-artists.deviantart.com/journal/8-Advantages-of-using-Linux-overWindows-291681914
[5] URL:]http://iverilog.icarus.com/
[6] URL: https://sites.google.com/site/iverilog/support/support-providers
[7] URL: http://opencores.org/opencores,tools

Simulating Cooperative Systems Applications: a New Complete Architecture

Dominique Gruyer
IM-LIVIC, IFSTTAR
Versailles, France

Sébastien Demmel
CARRS-Q, Queensland University of Technology
Kelvin Grove (QLD), Australia

Brigitte d'Andréa-Novel
CAOR – Centre de Robotique, MINES ParisTech
Paris, France

Grégoire S. Larue and Andry Rakotonirainy
CARRS-Q, Queensland University of Technology
Kelvin Grove (QLD), Australia

***Abstract*—For a decade, embedded driving assistance systems were mainly dedicated to the management of short time events (lane departure, collision avoidance, collision mitigation). Recently a great number of projects have been focused on cooperative embedded devices in order to extend environment perception. Handling an extended perception range is important in order to provide enough information for both path planning and co-pilot algorithms which need to anticipate events. To carry out such applications, simulation has been widely used. Simulation is efficient to estimate the benefits of Cooperative Systems (CS) based on Inter-Vehicular Communications (IVC). This paper presents a new and modular architecture built with the SiVIC simulator and the RTMaps™ multi-sensors prototyping platform. A set of improvements, implemented in SiVIC, are introduced in order to take into account IVC modelling and vehicles' control. These 2 aspects have been tuned with on-road measurements to improve the realism of the scenarios. The results obtained from a freeway emergency braking scenario are discussed.**

Keywords—Cooperative Systems; IEEE 802.11p; Inter-vehicular Communications; simulation

I. INTRODUCTION

Cooperative Systems (CS) are widely considered as the next major step in driving assistance systems (ADAS), aiming at increasing safety and comfort for drivers [1] Wireless Inter-Vehicular Communications (IVC) are used to share information so that drivers, or ADAS, can enhance their awareness of their surroundings. The state of the vehicle or the driver, detected objects and events pertaining to the driving environment (ranging from traffic and weather information to collision warning) are the type of information that can be exchanged within Vehicular Ad-hoc Networks (VANETs). A straightforward example of cooperative systems is Emergency Electronic Brake Light [2] (EEBL): a piece of information which is naturally available within a certain distance, i.e. a vehicle's break lights, is extended to a larger area of perception through IVC. Cooperative Collision Warning (CCW) can be achieved with EEBL by broadcasting a warning message whenever a vehicle is performing an emergency braking manoeuvre.

Development of CS requires additional resources in terms of extended perception which are both time-consuming and expensive. Therefore, it becomes essential to have a simulation environment or platform that allows prototyping and evaluating extended, enriched and cooperative ADAS in the early stages of the system's design. This virtual simulation platform has to integrate models of road environments, virtual on-vehicle sensors (proprioceptive & exteroceptive), infrastructure-based sensors and IVC devices, which are all consistent with the laws of physics. Similarly, a physics-based model for vehicular dynamics coupled with actuators (steering wheel angle, torques on each wheel) is required. Within such a platform, it becomes possible to simulate accurately the performance of future cooperative ADAS.

This paper presents an architecture to simulate and evaluate CS applications, based on the functionalities of both the SiVIC and RTMaps™ interconnected platforms [3], [4]. Such coupling allows meeting the aforementioned requirements. Our CS simulation architecture brings several improvements to the SiVIC-RTMaps™ coupling, regarding the modelling of IVC and vehicle's control.

The existing transponder-like behaviour of IVC simulation in SiVIC [5] is extended to a more realistic modelling with data from actual on-tracks measurements with prototype 802.11p devices. IEEE 802.11p [6] is the leading IVC technology that has been pushed forward by the IEEE for short-to-medium range communications (up to one kilometre), for both Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communications. To evaluate cooperative ADAS, especially when safety-critical tasks are concerned, it is necessary to be able to simulate 802.11p actual behaviour. Indeed, cooperative ADAS specifications and actual performance will be strongly affected by how 802.11p behaves on the road. Unfortunately, its performance is likely to diverge from that studied in earlier theoretical simulations, upon which most models are based, as we have shown in [7]. A safety-focused cooperative ADAS could have less benefit than we could expect in a real setting where IVC performance are overestimated.

Empirical modelling is a good way to improve simulations' accuracy by taking into account all existing perturbation sources. Thus, we have extended our transponder-like simulation so that it can have sufficient performance to emulate 802.11p. The new model outputs range, frame loss

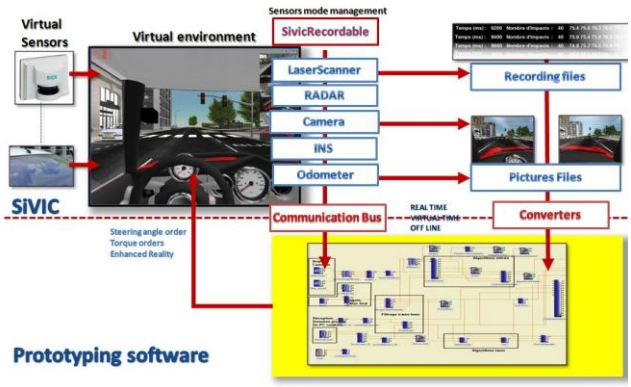


Fig. 1. General architecture of ADAS prototyping with SiVIC

and latencies, which are classified along the relative speed between vehicles and/or roadside units. Hardware

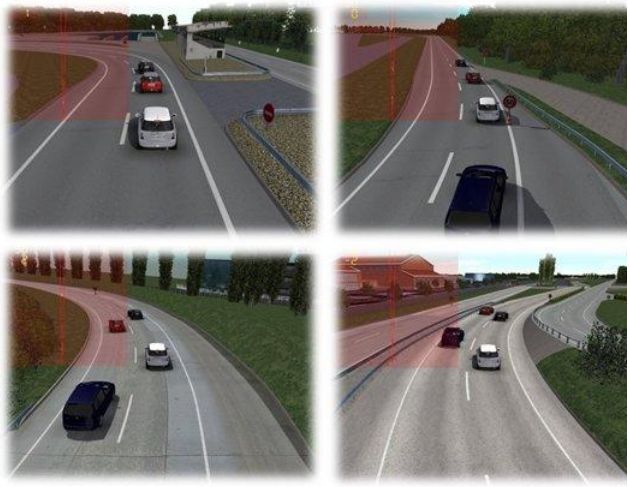


Fig. 2. Multiple captures from the Versailles-Satory's test tracks scenery; SiVIC's visual rendering is focused on generating acurate image dynamic, rather than merely acceptable realism to human eyes

inhomogeneities, ground reflections, multipath effects on vegetation and other objects, as well as the weather, are factors that influence the outputs. We have based our modelling on data collected on over 400 km of driving on Versailles-Satory's test tracks (near Paris, France) in late 2011 and early 2012.

Several improvements are also be made to the vehicle's controllers comparing to previous versions developed for the Full Range Speed Acc and Lane Keeping applications [8], in order to have a closer-to-life simulation of a human driver as well as introducing mechanisms related to CCW, such as emergency braking manoeuvres.

Our architecture can be used, for example, to evaluate the impact of introducing IVC devices into a driving situation leading to crashes, compared to using non-cooperative ADAS, or without any ADAS altogether. To demonstrate that our architecture can be used to produce meaningful results, we will show how an EEBL application can be simulated with it. In order to validate it, we have found that our application reproduces results from previous larger scale simulations [9], [10].

The remainder of this paper is organised as follows: Section II presents the CS simulation architecture we have developed, including software mechanisms in SiVIC and RTMaps™, the 802.11p IVC modelisation and our control's equations. Section III focuses on an application of our architecture, presenting detailed results analysing the effects a CS-based ADAS has on crash number and severity. Finally, we give few words of conclusion and perspective on future works in Section IV.

II. Cooperative Systems Simulation Architecture

A. SiVIC-RTMaps™ interconnection

Our CS simulation architecture is based on the interconnection of the sensors simulation platform SiVIC and the prototyping platform RTMaps™. The interconnection between SiVIC and RTMaps™ allows replacing real measurements by simulated ones, creating a fully Software-In-the-Loop (SIL) development and prototyping approach.

1) The SiVIC platform

SiVIC [3] is a platform designed to enhance the process of developing and evaluating ADAS. This platform enables the simulation of multi-frequency sensors embedded in static or dynamic devices, equipments and vehicles commonly used in ADAS scenarios.

The SiVIC platform is a very efficient tool to develop, prototype and evaluate high level ADAS (see Fig. 1), including CS applications. SiVIC can be easily interconnected with several external platforms such as RTMaps™ (see section II-A2) or Matlab™ (from Mathworks). This interconnection interface is efficient and useful to perform a great number of developments in a SIL approach. Once the application is evaluated in virtual condition and validated in simulation, it can be integrated and tested into a real embedded hardware architecture (on vehicle) further towards the end of the development cycle.

SiVIC uses a realistic graphical environment (Fig. 2), supported by physically accurate behaviours for vehicles and sensors. It can generate a flow of time-stamped and synchronised data that can be recorded and/or interacted with by prototyping and/or data treatment platforms such as RTMaps™ or Matlab™. Furthermore, SiVIC can generate multiple scenarios with events-driven mechanisms, so that the robustness and reliability of control and perception algorithms can be extensively tested on many parameters. This functionality is useful for scenarios featuring hazardous environments, complex situations, or missing or erroneous data (from sensors or actuators). Moreover, data analysis can always be performed with accurate ground truth references.

Proprioceptive and exteroceptive sensors are modelled in SiVIC so that, from the point of view of an algorithm, there is no difference between a fully SIL sensor and an on-vehicle sensor. Sensors available in SiVIC are:

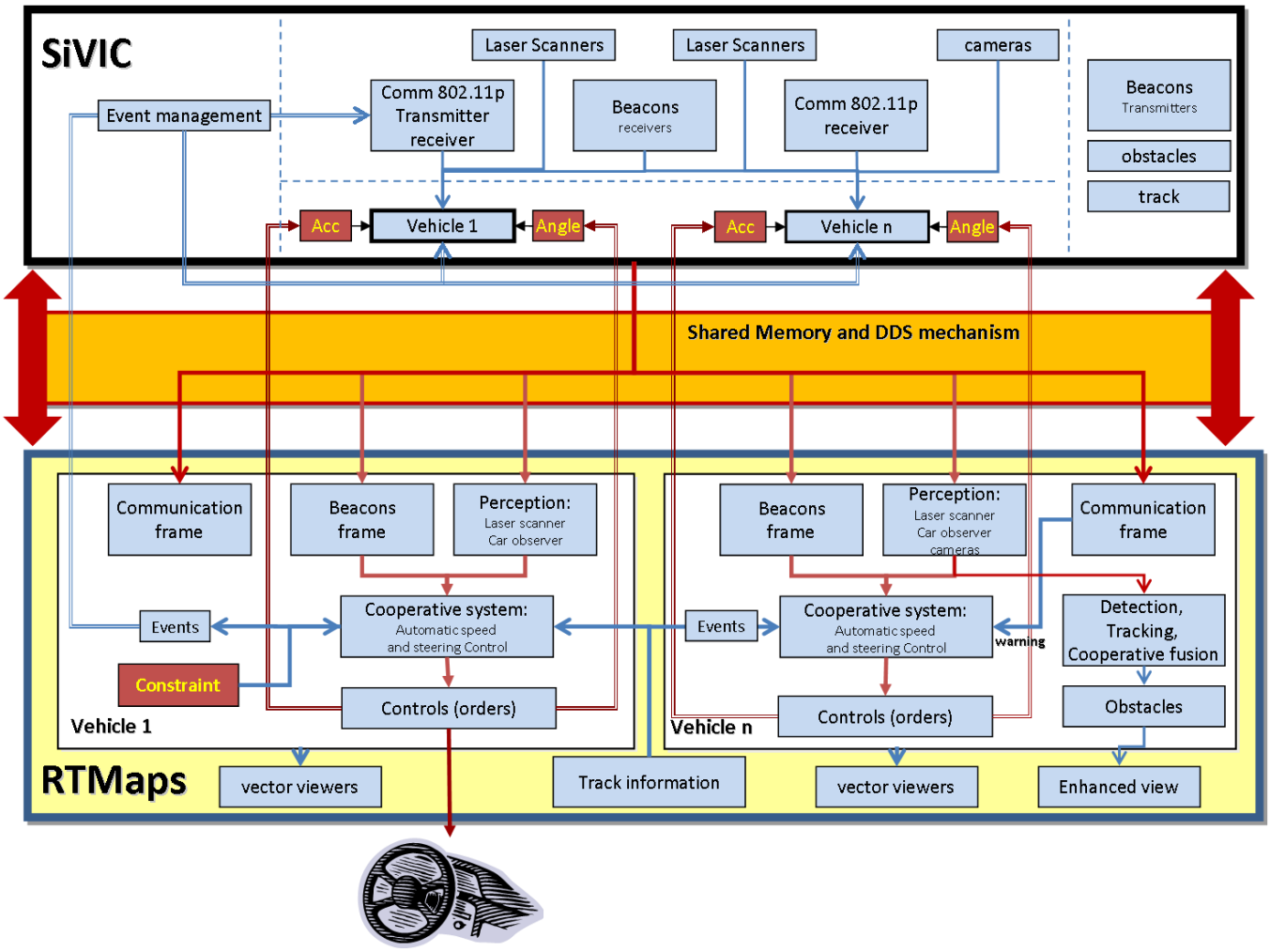


Fig. 3. CS simulation architecture's detailed functions in SiVIC-RTMaps™

- System variables (called observers) that provide output reference data on an object's position and behaviour
- Cameras (configurable either as software or hardware cameras), including Fisheye and omnidirectional cameras
- Inertial Navigation Systems (3 axes accelerometer + 3 axes gyrometer)
- Odometer
- Telemetric laserscanner (multi-layered, capable of using either ray-tracing or Z-Buffer methods)
- Radiofrequency transponders.

Additional sensors are being implemented in SiVIC at the moment and are close to deployment; most are related to the implementation of a realistic model of electromagnetic waves propagation in the platform. They include a simulation of GPS, that goes up to the influence of satellite's ephemerides, and a radar. The Radiofrequency transponders have been already used for previous works on cooperative speed control by transponder-equipped infrastructure [5]. This work has been used and extended in the present paper to take advantage of a more realistic model of IVC behaviour, based on field measurements.

2) RTMaps™

RTMaps™ is marketed by Intempora[1], based upon work undertaken at Mines ParisTech a decade ago [11]. Its primary goal is to record and process a large number of simultaneous data flows such as images, laserscanner scans, positioning data (from GPS, odometer or INS), etc. User-developed algorithms, for image processing or data fusion for example, can be deployed in the RTMaps™ framework in dedicated libraries called *packages*; packages themselves contain *components* to apply specific treatments to the data flow. Recorded data can be easily replayed, which is especially useful to precisely tune algorithms with multiple re-runs of a same on-tracks measurement.

[1] www.intempora.com

B. Overall architecture

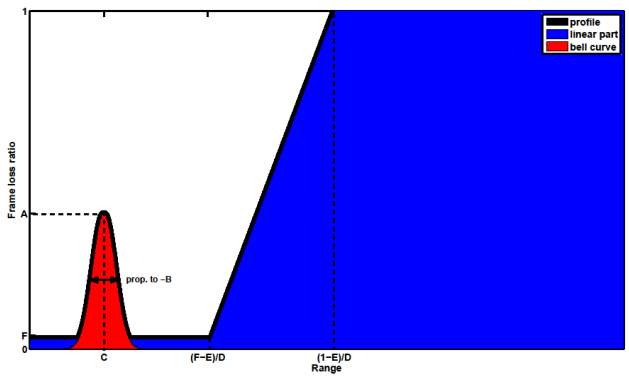


Fig. 4. Decomposition of a frame loss profile τ, with its parameters

The developed CS simulation architecture can be applied to any kind of CS usage scenario. It is partially scenario-specific, both in SiVIC and RTMaps. Scenarios are built in SiVIC by using script files, that specify the location of objects and many other parameters. The actual CS application needs to be developed and implemented in RTMaps. While developing our architecture, in order to simplify our work, we focused at first on simulating strings of IVC-equipped vehicles, specifically for EEBL applications. However, developments such as vehicles' control mechanisms and the 802.11p simulation are not scenario-dependent and are an integral part of SiVIC.

Fig. 3 shows the data flow and relationships between the environment, sensors and algorithms as implemented in the SiVIC-RTMaps™ environment. veh_1 is considered as the *leader*, and embeds a 802.11p modem used only for alert transmission, while the other vehicles (veh_n) are *followers*, and embed 802.11p only used in reception. Apart from this distinction, one can see that each vehicle has identical features. Vehicles are controlled from/within RTMaps, with the combination of the *cooperative system* and *control* blocks, that use information shared via IVC and obtained from embedded sensors to control the vehicle's behaviour in SiVIC (from actuators on the wheels and steering wheel). veh_1 can have additional control constraints, as required by the scenario.

C. Pre-existing transponders simulation

The pre-existing transponders simulation is composed of two plug-ins, for transmission and reception; both are strictly one-way. Each plug-in can be attached to any object in the simulation, from vehicles to roadside objects. The simulation is driven from the receptor's point of view. Whenever prompted, a receptor will check all the transmitters loaded in the simulation to verify whether they are in range, as specified in the receptor's parameters. Each emitter within range will then transmit its data frame. It is possible to use several receptors on a single vehicle, and receive messages from different transmitters.

The transponders were programmed to use the same frame pattern as real equipments. In order to match this format, the transponder uses three types of data: bits, bytes and strings, which are contained within a frame script. A frame script includes, for each line representing data, four fields: name, type, amount of data, and the actual data. Once these data are processed by the parser, a frame is produced. A receptor transponder decodes the frame pattern using the same algorithm as found in real devices.

D. 802.11p enhanced simulation

As initially used in [5], the transponders' range was very limited (less than 20 metres), as they only had to provide speed regulation information to passing vehicles. By extending the transponders' range, it was possible to get a simple simulation of IVC. However, this simulation was not representative of what would happen on a real road. Indeed, on the road, many factors will compound to alter the range of IVC and introduce errors that lead to frame loss, depending on the nature of objects inside the environment, weather, vehicles' behaviour, etc. Additionally, the existing transponders simulation did not include latencies, which, depending on the type of IVC used and the range of communication, can become non-negligible considering SiVIC's simulation step (fixed at 5 milliseconds in the current work).

Using data collected on the Versailles-Satory's test tracks during autumn and winter 2011-12 [7], we developed a new, more accurate 802.11p-based IVC model that introduces realistic IVC defects. The model covers the following sources of defects: multi-path reflections from the ground, from objects and vegetation, from weather, and from hardware inhomogeneities. Although they are accounted for in the model, these sources of performance variations cannot be differentiated, nor can their relative importance be modified. As such, the model is representative of the conditions on Satory's tracks, that we assume to be fairly similar to those found on most freeways.

This model provides a frame loss probability depending of the distance and relative speed between two nodes (vehicles or RSUs); more precisely the model generates a *semi-linear frame loss profile*. A profile represents a single uninterrupted connection between two IVC devices and is used to determine the frame loss probability at any given distance, as long as they are within range (that is, as long as frame loss is under 100\%). By using profiles, we focus on modelling individual instances of connection between two nodes, rather than simply modelling the average frame loss obtained over our total experimental measurements. A generic profile is shown in Fig. 4.

A profile is actually composed of two parts: a bell-shaped curve at first, followed by a linear regression; each part representing a different phenomenon. A profile τ is described by equation 1:

$$\tau = \max\left[A.e^{B.(d-C)^2}\,;\min\left(\max\left[D.d+E\,;F\right];1\right)\right] \quad (1)$$

where d is the distance between the emitter and receptor. $A,B,\ldots,F$ are the model parameters estimated from empirical data.

The term $A.e^{B.(d-C)^2}$ represents the frame loss area corresponding to the strongest ground reflection interferences,

centred at distance *C*. At this point the ground-reflected signal is strong enough to cancel out a large proportion of the incoming direct signal's energy, pushing a proportion of frames under the chipset reception's threshold; the frame loss corresponding to this proportion is represented by *A*. The bell curve's width is proportional to *B*; note that *B* is always negative. The model assumes that no counter-measure is applied to reduce the frame loss induced by interferences at *C*.

The term $D.d + E$ is a linear regression where τ is modelled linearly as a function of distance *d* and parameters *D* and *E*. This term represents the progressive increase of frame loss as received signal strength decrease. The increase starts from a non-zero frame loss ratio value given by parameter *F*, which represents the average of small perturbations measured within range. Typically, *F* will be low (less than 5%). *D* and *E* have two meaningful ratios: ratio $\frac{F-E}{D}$ gives the distance at which frame loss starts to increase from the plateau at *F*; ratio $\frac{1-E}{D}$ expresses the distance at which frame loss reaches 100%, hence the maximum range.

We then created four classes, which are classified according to the relative speed between the emitter and receptor. The classes are:

$$\text{speed} = [0; 40],]40; 60],]60; 100],]100; 160]$$

The first speed interval is for equivalent speed between emitter and receptor. The last speed interval is dedicated to the opposite traffic direction, or for a scenario with one static actor and another dynamic one. The 2 last intervals represent other cases (acceleration, deceleration, overtaking, etc.).

For each class, the model's parameters *A,B,...,F* are estimated using the Levenberg-Marquardt algorithm for non-linear least squares [12]. Experimental data show that *D* and *E* are linearly correlated; the other parameters are assumed to be independent. The relationship between *D* and *E* is given by a Generalised Linear Model regression from the observed values of *D* and *E*:

$$E = \alpha D + \beta + e, \quad e \rightsquigarrow \mathcal{N}(0, \sigma) \tag{2}$$

For each parameter of the vector $(A, B, C, D, F)^T$ applied to a specific class, a non-parametric probability density estimate is computed. The continuous distributions **A, B, C, D, F** of each parameter of the vector $(A, B, C, D, F)^T$ are computed with a Gaussian kernel smoothing method (the distribution **E** of the parameter *E* can be obtained through its linear correlation with **D**).

The transponders' functions described in the previous subsection are kept with our new approach. However, when a receptor checks whether it is in range with transmitters, new tests are applied. At first, a frame loss profile is generated from parameters distributions selected from the appropriate class for the relative speed between the transmitter and the receptor. To reduce computational load, a new profile is generated only when the distance is under a certain threshold (typically, 1,000 metres).

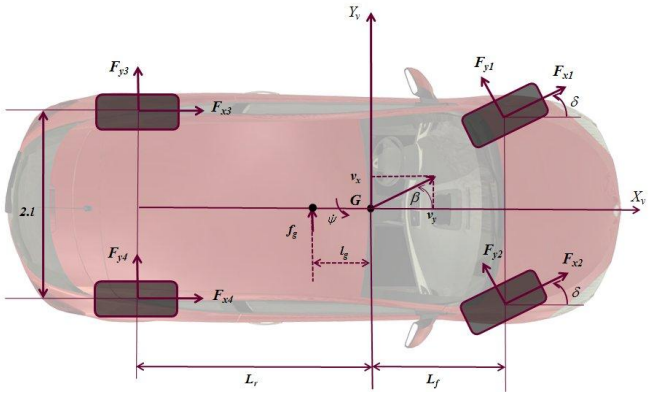


Fig. 5. Dynamic vehicle model

If a connection was previously established and lost for more than a certain duration (at least 30 seconds), a new profile is generated too.

After the profile is generated, the frame loss probability is extracted and the frame's success is tested against this value. In case of success, the receptor is allowed to read the frame's content according to the existing procedure. In the mean time, the profile is tagged as ``active'' and continues to be used for any frame exchange between these two specific transponders. If *n* is the number of transponders in the simulation, the maximum number of active profiles is thus $\frac{n(n-1)}{2}$.

Latency can be applied at this stage, by delaying the frame's extraction by a number of simulation steps. However, we have shown in [7] that point-to-point latencies remained overwhelmingly (99.47%) under 4 milliseconds for frames less than a 500 bytes (typical for EEBL applications). A simulation step is 5 milliseconds in SiVIC's default configuration, which means that short frame can be passed to the receptor transponder without delay. Nevertheless, if the simulation time is fixed to a lower value, then the latency mechanism could be activated.

We implemented a test mechanism based upon the amount of data which is encoded into the frame: if the amount is larger than a threshold (500 bytes), a delay is applied on the frame's data extraction. The number of simulation steps by which the frame is delayed is based on a simple linear regression from latencies measured experimentally. For example, a 500 bytes frame would be delayed by one step, which will provide a total latency of 10 milliseconds.

E. Vehicles' control

SiVIC provides a parametric model developed by Sébastien Glaser [13] (see also [14]) for the dynamic behaviour of the vehicle chassis on the three axes (roll, pitch and yaw/heading). It also accounts for shock absorbers dynamics and non-linear tire road forces [15], [16]. Eventually, coupling between longitudinal and lateral axes, the impact of normal force variations, and the car alignment's moment are also integrated. The vehicle's chassis is modelled with an unbending suspended mass. This model allows installing, in a simple way, a large number of on-board sensors. We will use the notations of the chassis dynamics illustrated in Fig. 5.

In order to obtain the most accurate sensor data, comparatively to a real situation, it is necessary to both handle a vehicle dynamical model, and to simulate realistic actuator models. The actuators are motor and braking torques applied on each wheel, and the steering wheel angle. One can thus

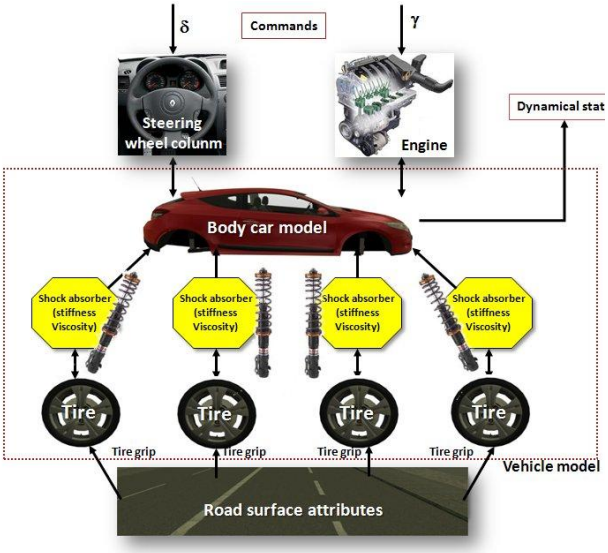


Fig. 6. Vehicle model in the SiVIC platform with its attributes

simulate front wheel drive, rear wheel drive or four wheels drive.

Figure 6 shows the model's level of complexity, and the links between all the different physical modules involved in it. Each module is completed with a list of available parameters. These parameters can be modified during the simulation. Vehicles' control is based on the same architecture as described in [5]. Vehicles are controlled longitudinally by torques on the wheels and laterally by the steering wheel angle; controls are decoupled.

Similarly to the previous architecture, lateral control is performed with an accurate map of the test track: angular and lateral deviations from the vehicle's lane are computed from this map. The controller uses the road's curvature at the vehicle's position, the inter-axles distance, and the angular and lateral deviations as described by equation (3):

$$\delta(t) = \tan^{-1}\big(L \times K(t)\big) \quad (3)$$

with

$$K(t) = K_{ref}(t) - [\mu_{\perp} \times (\psi - \psi_{\text{ref}}) + \lambda_{\perp} \times e_{\perp}] \quad (4)$$

In the previous lateral equations (3) and (4), $\delta(t)$ is the lateral steering angle, L is the inter-axle distance, set at 2.58 metres. $K(t)$ is the correction term on the vehicle's curvature, depending on the road's curvature. This correction depends of two suitable gains $\mu_{\perp}$ and $\lambda_{\perp}$. ψ and ψ_{ref} are, respectively, the vehicle's yaw angle and the road's heading. $e_{\perp}$ is the lateral deviation. If $\delta(t)$ is greater than δ_{max}, then we apply a saturation stage:

$$\delta(t) = \frac{\delta(t)}{|\delta(t)|} \times \delta_{max} \quad (5)$$

In our application, vehicles can be asked to follow the left, central or right lane during the simulation. If required, lane detection and tracking can be used instead of a track map, so that any simulated road can be used.

Longitudinal control has been improved from the previous architecture. Previously, vehicles were simply instructed to follow a certain speed, which was modified manually or from roadside beacons using the transponders simulation. This mechanism is kept, although it is now overridden by two additional controls.

Firstly, a mechanism is added in order to simulate an *interdistance regulation* process. As our typical demonstration scenario involves a platoon of several vehicles following each others, vehicles need to remain within acceptable interdistances at all times. On each vehicle, a pitch-stabilised narrow-beamed laserscanner is used to measure the distance to the leading vehicle. To maintain an acceptable interdistance, the vehicle's reference speed (or speed target) V_{ref} is computed with equation (6).

$$V_{ref} = V - \big[V.(t_{inter} - t_h) - d_{target}\big] \times \frac{1}{t_{inter} - t_h} \quad (6)$$

where V is the vehicle's current speed, t_{inter} the minimum acceptable intervehicular time, t_h the driver's reaction time and d_{target} the distance to the closest obstacle, as measured by the laserscanner. This mechanism is used to simulate a simple human driver behaviour and the intervehicular time respected by the driver.

For the leader vehicle, the reference speed is extracted from frames received from the infrastructure transponders. When a receiver attached to the leader vehicle receives the new speed information, the following control is applied:

$$Ct = 3 \times R \times M \times \big(V - V_{ref}\big) \quad (7)$$

where Ct is the torque order applied to the front wheels, R the wheel's radius, and M the chassis' mass. V is the leader vehicle's speed and V_{ref} is the reference speed. For a follower vehicle, the same equation is used but with V_{ref} computed from equation (6).

A second approach has been developed in order to maintain a Time To Collision (TTC) around 2 seconds. From a speed V_f (follower vehicle's speed), the distance required to maintain the 2 seconds TTC is $D(t) = V_f(t) \times 2.0$. Then, the safety distance is $e = D_{lf}(t) - D(t)$, where D_{lf} is the vehicular interdistance between a leading vehicle and its follower. The "safety speed" $\dot{e}$ is also estimated. From there, the control applied to the wheels is computed as follows:

$$Ct = M \times R \times (Kp \times e + Kd \times \dot{e}) + \sum_{i=1}^{4} I_i \times \dot{\omega}_i \qquad (8)$$

With $\dot{\omega}_i$ the derivative speed of wheel *i*, *Kd* the derivative gain and *Kp* the proportional gain. Suitable values have been set for *Kd* and *Kp*. If *Ct* is negative, then the current manoeuvre is a deceleration, and *Ct* is applied to the four wheels (*Ct/4*). If *Ct* is positive, then the current manoeuvre is an acceleration, and the torque order is applied only to the front wheels (*Ct/2*).

Secondly, we have an *emergency regulation* mechanism. This mechanism is triggered only on IVC-equipped vehicle, when an emergency braking frame is successfully received and decoded by the receptor. Immediate or delayed reaction can be chosen, allowing simulating either a reactive or informative system (i.e. one with automated braking versus one that simply flashes an alert to the driver). In the former, V_{ref} is simply set to zero immediately after the frame is decoded. In the latter case, V_{ref} is only updated after a t_h delay has passed. At the moment, the only way for the vehicle to not brake is to miss the emergency braking frame. A future extension will allow a more realistic behavior with a context-aware, so that vehicles which are far away from the actual event (e.g. more than 500 metres) and still receive an emergency braking frame either ignore it, or enter into a state of heightened alert (where t_h is decreased and t_{inter} increased). The leader vehicle has a similar mechanism for the initial emergency braking, which is triggered when its curvilinear abscissa on the tracks reaches a user-defined value.

III. Cooperative Collision Warning Prototyping

We implemented an EEBL/CCW application with our architecture, which was inspired from the scenario studied in [17], [9], [10]. Results from [9] showed that only a small percentage of IVC-equipped vehicles was necessary in a vehicles platoon to considerably reduce the number of crashes, which was confirmed in [10]. For example, in dense configuration, only 5% of equipped vehicles were sufficient to reduce the number of crashes by two thirds in an emergency braking scenario compared to completely unequipped scenario. We will show how our architecture can reproduce these previous results (on a smaller scale) and show that they can be refined when a more detailed simulation architecture is available.

The previous studies used heavily constrained strings or platoons of vehicles. However, the interest of this architecture is its capacity to generate generic and non repeatable configurations in order to be closer from reality. Thus, we will use a scenario which is less constrained and non-repeatable. That way, we will be able to compare our results and previous studies, and test whether the results from previous studies still hold when the vehicles platoon behaved more realistically. As we will see later on, this is not completely the case.

We set up a scenario which is identical in practice to the scenario studied in [9], [10], with the only difference being the reduced size of the vehicles' platoon. The granularity of our results will be limited compared to studies using larger strings, but it should not be an impairment to the validation of our CS simulation architecture.

A five vehicles platoon (1 leader, 4 followers) is set up in SiVIC, in the virtual reproduction of Versailles-Satory's test track called *la routière*, modelling a French non segregated trunk road (*route nationale*). Each vehicle can be configured individually and independently, but for the sake of simplicity, we will keep an homogeneous fleet in terms of acceleration, braking capacity and reaction time ($t_h = 0.5$ second). The same was true in [9], [10]. Additionally, the vehicles are set in reactive mode: there is no delay between the reception of an emergency frame and the beginning of the braking action. All vehicles have $t_{inter} = 2.5$ seconds; except veh_2 for which $t_{inter} = 1.5$ seconds, in order to simulate a slightly more risked driving. According to government statistics, more than half (56.4%) of the drivers do not follow safe interdistances recommendations (at least 2 seconds) in dense traffic [18]. All these parameters are either controlled from RTMaps platform (where they can be changed online), or in the SiVIC script, which is loaded once at start-up.

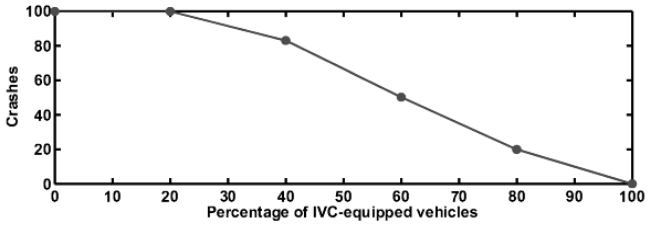
The vehicles start with a static configuration and all grouped together in one location of the track. From these starting positions, the vehicles arrange themselves in a platoon on the right-hand lane, and progressively speed up to 70 km/h. While the starting positions are always identical, the interdistance regulation at very short distances means that at each instances there are varying interdistances between the five vehicles; each scenario's instance forms a different platoon. Follower vehicles are equipped with telecommunication receptors, depending on the desired equipment ratio. IVC equipment is randomly selected for each individual follower. The equipment is reset at each new run. The emergency braking event takes place in a long straight section approximately 700 metres after the starting position.

The scenario was replayed at least one hundred times for each of the following equipment ratios: $0/5$, $2/5$ (leader + 1 follower), $3/5$, $4/5$, and $5/5$. The $\rho = 1/5$ case is not simulated as it corresponds to having only the leader vehicle equipped, which is not different from $\rho = 0/5$ in this scenario. A total of 716 runs were simulated, which generated 1197 crashes. The following variables were recorded for all vehicles: curvilinear abscissa, TTC (Time To Collision), the distance from the obstacle d_{target}, the ego-vehicle speed *V*, the speed reference V_{ref}, emergency frame broadcast and instances of collisions.

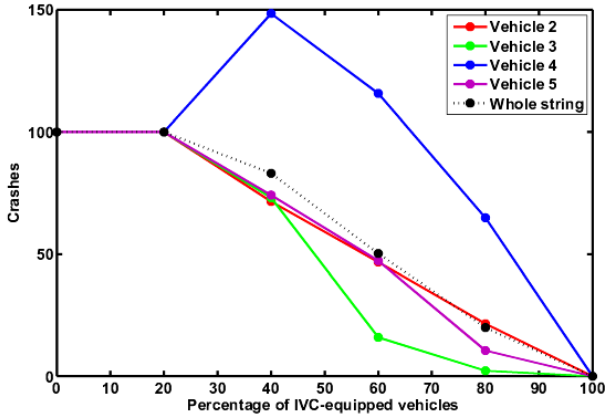
1) Crashes number analysis

Fig. 7a shows the normalised total rear-end crashes at different equipment ratios. By introducing 2 IVC-equipped vehicles ($\rho = 2/5$, or 40%), the number of crashes fell by 17%; with $\rho = 3/5$, the crashes fell by 50%, and with $\rho = 4/5$, the crashes fell by 80%. In a completely equipped platoon, no crashes were recorded. Note that the crashes number is maintained at 100\% for $\rho = 1/5$ since it is indistinguishable from $\rho = 0/5$.

As we already stated, in [9], [10] the number of vehicles in the string was significantly higher, which allowed for a better granularity of ρ. Compared to the repeatable scenario however, the large number of simulation runs make it possible to obtain more refined, and more realistic, results. Contrary to [10], our results do not show a strong $1/x$ type decrease of



(a) Normalised crashes count for the whole string



(b) Normalised crashes count for the whole string and each individual vehicle

Fig. 7. Illustrations of the reduction in crahses obtained by introducing IVC in the vehicles string

crashes when the IVC equipment ratio increases. However, they follow the same trend; for example, at a 2,600 vehicles/hour capacity, the reduction in crashes' number from $\rho = 0\%$ to $\rho = 80\%$ is very similar to our results. Furthermore, it can be noted that in our scenario, IVC equipment starts to provide a reasonable safety increase with only more than 50% of equipped vehicles. This difference can

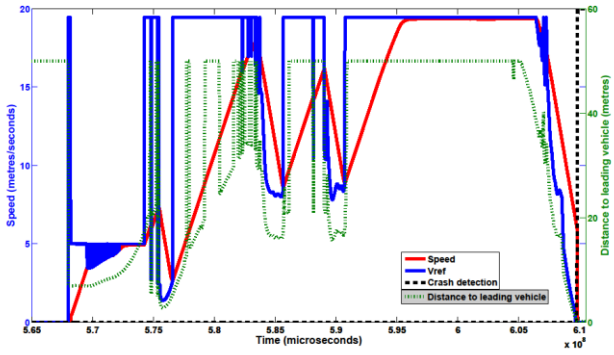


Fig. 8. Detailed variables measurements for one vehicle during a simulation run

be attributed to the different methodologies applied to the studies. Nonetheless, our results are coherent with the EEBL scenario and shows that our CS simulation architecture can be used to complement larger simulations like [10], notably by providing more detailed analysis. Indeed, being able to record and study variability inside the platoon, for each individual vehicle, is a major improvement brought by the architecture. Different kinds of data can be considered for study, which will be shown with two examples. Typically, it is possible to extract information on the behaviours of each vehicle. This can concern the behaviour of a vehicle during a single run (our first example), or crash patterns associated with them over the whole experimental runs (our second example).

On one hand, we can focus on a single vehicle behaviour. For example, Fig. 8 shows the interdistance regulation variables for veh_3, taken during a $\rho = 3/5$ run. In this run the veh_3 was not equipped with an IVC device. V_{ref} is shown by the blue curve (left-hand axis). It depends either on limit speed instructions from RST (Road Side Transponder) or on d_{target}, shown by the green dotted curve (right-hand axis), via equation (6). The quick distance variations visible on this figure usually happen when the preceding vehicle is turning, and so leaves the narrow field of view of the frontal laserscanner. One can also note, just after the 5.7×10^8 timestamp, a brief period for which V_{ref} fluctuates a lot, very quickly. This is a visual representation of the interdistance regulation behaviour that leads to the scenario's non-repeatability.

V (red curve, left-hand axis) is well regulated according to V_{ref}. At the end of the run, we can note that the vehicle starts to brake because d_{target} becomes too small; the delay introduced by the human reaction time is clearly visible: when the vehicle starts to brake, the interdistance has already shrunk by more than 10 metres. Even at maximum braking power, veh_3 cannot stop before the impact with the preceding vehicle, which is shown by the vertical black dotted line. The impact takes place at the relatively slow speed of 6 metres.seconds^{-1}(~22km/h).

On the other hand, we can focus on whole runs. Fig. 7b reproduces Fig. 7a data (the dashed black curve), and overlays it with the normalised crashes counts for each individual vehicle. From this figure, it is easy to see that when the IVC equipement grows then the number of crashes decreases. However, it seems not to be necessarily the case for individual vehicles. Indeed, for veh_4 and the scenario having 40% of IVC equipment, we observe a 47% increase in the number of crashes encountered by this vehicle. At 60% equipment ($\rho = 3/5$), the crash count is still 15% higher than in a fully non-equipped scenario. At 80% equipment, veh_4 benefited from IVC the same way that other vehicles benefited for 40% equipment.

Taken at face value, this result would suggest that while drivers would collectively benefit from using EEBL, some drivers unfortunately would see their crash likelihood increase. Obviously, this is an unacceptable conclusion in terms of road safety. It is further aggravated knowing that, if the absolute number of crashes is considered, veh_4 is the one with the least crashes at $\rho = 0/5$. How could introducing IVC make the previously safest vehicle the least safe? Further investigations show this is unlikely to happen in an actual on-road situation.

Indeed, from the detailed recordings, veh_4 appears to be following veh_3 with an interdistance slightly above average. In the scenario without IVC, this is not an issue and the vehicle manages to stop before colliding with veh_3 in most cases,

hence its lower number of crashes relatively to the others. On the other hand, at $\rho = 2/5$, if veh_2 or veh_3 are equipped with IVC, veh_3 will have a tendency to brake earlier than it did in the non-equipped scenario. Because of this, the relative speed between the two vehicles is large when veh_4's controller starts to brake. In that case, even at maximum braking capability, veh_4 is unlikely to be able to stop before colliding with veh_3, which leads to the 50% increase in crashes we measured.

Unfortunately, this behaviour stands out as a limitation of our simulation. Indeed, on a real road, veh_4's driver would become aware of veh_3's braking manoeuvre with the activation of its braking lights; veh_4 would thus brake much earlier than what the current controller decides to do. At the moment, our architecture cannot simulate this behaviour. Our scenario is thus artificially increasing crashes for that specific vehicle.

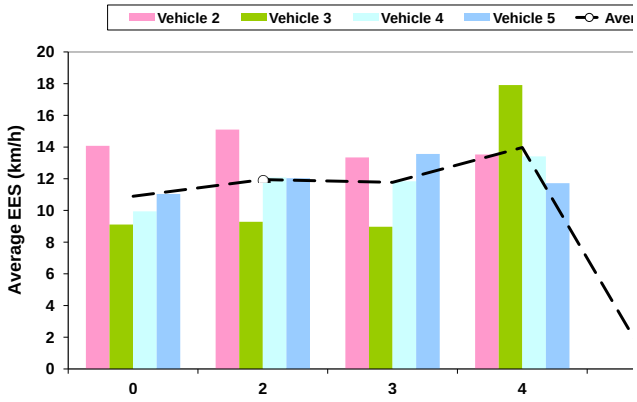


Fig. 9. Average EES computed for each and all vehicles, at different values of ρ

Nonetheless, this happens only with veh_4, which do not invalid the results for the whole platoon. All the remaining vehicles behaved according to the scenario's expectations. Additionally, if we filter the runs to keep only the ones where veh_4's interdistance is comparable to the string's average, veh_4 behaves like the other vehicles. Also note that veh_2's count is higher than the others on average (ignoring veh_4) because of its more aggressive driving style, which is consistent with the scenario's setting.

2) Crashes severity analysis

Thanks to SiVIC's realistic vehicle motion models, we can estimate the severity of crashes from the EES (Equivalent Energy Speed), which is the energy dissipated by the velocity change when a vehicle is hitting an obstacle. This analysis is made in post-processing based on the vehicle's variables recorded during the simulation runs.

Interestingly, the EES results (Fig. 9) show that while increasing IVC equipment leads to less crashes, it does not reduce the remaining crashes' severity, except for complete equipment where no crash took place. The dispersion of individual averages does not allow concluding that severity actually increased. However, the severity is demonstrably not decreasing, contrary to what was found in previous studies (unless of course when $\rho = 5/5$ where there is no crash). A look into the detailed distribution of EES for each individual vehicle confirms this lack of improvement. The shapes of the EES distribution fluctuate, but the averages remain relatively stable or can even increase in some case

Note that the veh_3 outlier (94% increase) at $\rho = 4/5$ is computed from only two crashes on 224 runs. In the two runs where it crashed, veh_3 was following the preceding vehicle very closely to the minimum acceptable interdistance, and thus did not have the time to react properly during the emergency braking event. If the standard deviation is small for this vehicle, it is because the two crashes took place in runs that happened, by chance, to be almost exact repetitions.

While the EES absolute values remained largely under any dangerous threshold due to the scenario's conditions, implications are worrying at higher speeds. Indeed, from the point of view of a system's contribution to road safety, it is better to have several weak crashes, where no driver is injured, then one or two violent ones, where there are fatalities. In [9], it was shown that using the raw crashes number to evaluate IVC's contribution to the platoon's safety was always more pessimistic than using an EES-based severity criterion. However, we found here that while the number of crashes indeed significantly decreased, the remaining crashes' severity did not decrease. In this case, a crashes number-based criterion would have been considerably more optimistic than the EES-based severity criterion.

IV. Conclusions and Future Works

In this paper we have presented a cooperative systems simulation architecture, developed within the SiVIC-RTMaps™ interconnected platforms. This architecture uses the SiVIC-RTMaps™' capabilities to provide very realistic simulations, and has several improvements on previous architectures developed at LIVIC. The two main improvements concern: (1) firstly, the introduction of an empirical modeling of 802.11p IVC system based on ground-truth data collected on the Satory's test tracks; and (2) secondly, an improved vehicle controller, allowing for an automated vehicle to behave more like a human-driven one. The many variables accessible in great details, supported by realistic physical models (e.g. for vehicle's motion), also provide an improvement on pre-existing simulations, so that the behaviour of individual vehicles can be studied.

We validated this architecture by reproducing results from previous researches on the contribution of IVC to the reduction of rear-end crashes in vehicle platoons, with a caveat. Compared to these previous results, our architecture allows diving into greater details into each vehicle's behaviour, as many different variables are accurately recorded. Individual statistics can be generated for each vehicle. We have used these functionalities to evaluate the severity of each individual crash (a total of 1197, over 716 runs of a 5-vehicles string), via the computation of the Equivalent Energy Speed (EES). This more detailed study has yielded unexpected results: while introducing IVC decreases the number of crashes as expected, the average EES does not decrease. This means that the remaining crashes' severity remain constant. These results need to be further confirmed, in which case they would raise a few concerns about the actual safety benefits of IVC. It is often assumed that IVC will also help to reduce the severity of crashes, and that in some cases it might be more beneficial to

favour this effect rather than simply reducing crashes numbers. If some benefits can be expected from the reduction of crashes that we obtained, our results show that the efficiency of an EEBL application to reduce crashes severity might have been over-estimated. Effectively, the safety benefits of IVC for road users are not as important as initially expected. Earlier results suggested that introducing IVC would lead to less crashes, and that the remaining crashes would be less severe. Our results suggest that, indeed, there will be fewer crashes. However, the remaining crashes will remain as severe as previously.

Further work should concern determining whether the absence of improvement of severity is a by-product of our scenario's setting, and continuing on improving the architecture's functionalities. A new control mechanism is required to better reproduce drivers' behaviour when the emergency event is taking place more than a few dozen metres in front of them. Related to this issue, vehicles currently react immediately to the reception of an emergency braking frame, both in reactive or informative modes, which just modulate the reaction's delay. Thus, equipped vehicles located several hundred metres away from the initial perturbation will also start to brake, when that is, in most cases, not necessary. In further studies, we will limit the immediate reaction to a certain radius around the initial perturbation. Vehicles outside this radius will be put into a state of heightened alert, by increasing t_{inter} and decreasing t_h. Moreover, we will test the impact of degraded vehicle dynamic conditions like low tires adherence, low braking capacities, strong acceleration capacities, etc. in such an IVC safety systems.

ACKNOWLEDGEMENT

This work is supported by the Commonwealth of Australia through the Cooperative Research Centre for Advanced Automotive Technology, as well as by the French Institute of Science and Technology for Transport, Development and Networks. A sub part of this work has been developed in the CooPerCom project, a 3-year international research project (Canada-France). The authors would like to thank the National Science and Engineering Research Council (NSERC) of Canada and the Agence nationale de la recherche (ANR) in France for supporting the project.

REFERENCES

[1] S. Demmel, D. Gruyer, and A. Rakotonirainy, "V2V/V2I augmented maps : state-of-the-art and contribution to real-time crash risk assessment," in 20th Canadian Multidisciplinary Road Safety Conference. Hilton Niagara Falls, Ontario: The Canadian Association of Road Safety Professionals, June 2010.

[2] A. A. Carter, "The status of vehicle-to-vehicle communication as a means of improving crash prevention performance," National Highway Traffic Safety Administration, Tech. Rep, pp. 01–19, 2005.

[3] D. Gruyer, S. Glaser, and B. Monnier, "SiVIC, a virtual platform for ADAS and PADAS prototyping, test and evaluation," in FISITA World Automotive Congress, June 2010.

[4] D. Gruyer, S. Glaser, S. Pechberti, R. Gallen, and N. Hautiere, "Distributed simulation architecture for the design of cooperative ADAS," in First International Symposium on Future Active Safety Technology toward zero-traffic-accident, September 2011.

[5] D. Gruyer, S. Glaser, B. Vanholme, and B. Monnier, "Simulation of automatic vehicle speed control by transponder-equipped infrastructure," in 9th International Conference on Intelligent Transport Systems Telecommunications, October 2009, pp. 628–633.

[6] IEEE Computer Society, "IEEE standard for information technology - telecommunications and information exchange between systems – local and metropolitan area networks - specific requirements part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications - amendment 6: Wireless access in vehicular environments," New York, 2010.

[7] S. Demmel, A. Lambert, D. Gruyer, A. Rakotonirainy, and E. Monacelli, "Empirical IEEE 802.11p performances evaluation on test tracks," in IEEE Intelligent Vehicles Symposium, 2012.

[8] D. Gruyer, S. Pechberti, and S. Glaser, "Development of full speed range acc with sivic, a virtual platforme for adas prototyping, test and evaluation ADAS," in IEEE Intelligent Vehicles workshop, june 2013.

[9] B. Mourllion, D. Gruyer, and A. Lambert, "A study on the safetycapacity tradeoff improvement by warning communications," in IEEE Intelligent Transportation Systems Conference, September 2006, pp. 993–999.

[10] A. Lambert, D. Gruyer, A. Busson, and H. M. Ali, "Usefulness of collision warning inter-vehicular system," International Journal of Vehicle Safety, vol. 5, no. 1, pp. 60–74, 2010.

[11] B. Steux, "RTMaps, un environnement logiciel dédié à la conception d'applications embarquées temps-réel. utilisation pour la détection automatique de véhicules par fusion radar/vision," Ph.D. dissertation, Écoles des Mines de Paris, 2001.

[12] J. Moré, "The Levenberg-Marquardt algorithm: Implementation and theory," in Numerical Analysis, ser. Lecture Notes in Mathematics, G. Watson, Ed. Springer Berlin / Heidelberg, 1978, vol. 630, pp. 105–116.

[13] S. Glaser, "Mod´elisation et analyse d'un v´ehicule en trajectoires limites. application au d´eveloppement d'un syst`eme d'aide `a la conduite," Ph.D. dissertation, Universit´e d'Evry Val d'Essonne, 2004.

[14] N. Kiencke and L. Nielsen, Automotive Control Systems: For Engine, Driveline and Vehicle. Springer Verlag, 2004.

[15] H. Dugoff, P. Fancher, and L. Segel, "An analysis of tire traction properties and their influence on vehicle dynamic performance," SAE Technical Paper, Tech. Rep., 1970.

[16] R. Rajamani, Vehicle Dynamics and Control. Springer USA, 2006.

[17] M. Mangeas, "Rapport d'activit´es, livic," IFSTTAR (LIVIC), Tech. Rep., 2003.

[18] Observatoire national interministériel de la sécurité routière, "La sécurié routière en france - bilan de l'année 2010," Paris, 2011.

Construction of Powerful Online Search Expert System Based on Semantic Web

Yasser A. Nada
Chairman of Department of computer science,
Faculty of Computers and Information Technology,
Taif University - K S A,
y_nada@yahoo.com

Abstract— In this paper we intends to build an expert system based on semantic web for online search using XML, to help users to find the desired software, and read about its features and specifications. The expert system saves user's time and effort of web searching or buying software from available libraries. Building online search expert system is ideal for capturing support knowledge to produce interactive on-line systems that provide searching details, situation-specific advice exactly like setting a session with an expert. Any person can access this interactive system from his web browser and get some questions answer in addition to precise advice which was provided by an expert. The system can provide some troubleshooting diagnose, find the right products; … Etc.

The proposed system further combines aspects of three research topics (Semantic Web, Expert System and XML). Semantic web Ontology will be considered as a set of directed graphs where each node represents an item and the edges denote a term which is related to another term. Organizations can now optimize their most valuable expert knowledge through powerful interactive Web-enabled knowledge automation expert system. Online sessions emulate a conversation with a human expert asking focused questions and producing customized recommendations and advice. Hence, the main powerful point of the proposed expert system is that the skills of any domain expert will be available to everyone.

Keywords— Expert System, Semantic Web, Online Search, XML.

I. INTRODUCTION

The Semantic Web is an extension of the World Wide Web with new technologies and standards that enable interpretation and processing of data and useful information for extraction by a computer. The World Wide Web Consortium (W3C) recommends XML, XML Schema, RDF, RDF Schema and Web Ontology Language (OWL) as standards and tools for the implementation of the Semantic Web. Ontologies work as the main component in knowledge representation for the Semantic Web. It is a data model that represents a set of concepts and the relationships between those concepts within a domain. Building an ontology starting from scratch is not an easy task since it makes heavy demands on time in addition to expert knowledge related to the domain.

Expert System (ES), also called a Knowledge Based System (KBS), is computer application programs that take the knowledge of one or more human experts in a field and computerize it so that it is readily available for use. It can also be integrated with textual database which can be used for explanation purposes of basic terms and operations to confirm and to reach conclusion in some situations [1].

A challenge for the Semantic Web is enabling information interoperability between related but heterogeneous ontology [2].

One of the most powerful attributes of expert systems is the ability to explain reasoning. Since the system remembers its logical chain of reasoning, a user may ask for an explanation of a recommendation and the system will display the factors it considered in providing a particular recommendation. This attribute enhances user confidence in the recommendation and acceptance of the expert system [3].

Each time we require certain software or product, we searched the internet twice and more, and maybe 10 times to find what we are searching for, that waste our efforts and time. Many times, we did not even found the software, and then we have to go and buy it from the CD's Shop. Also software fall in hundreds of categories, and sometimes we are very confused what to use and what is the exact features of each one of them. We search through the website of that software and find that all the data is crowded and we just can't differentiate between a software and another and we got confused from that.

That paper intends to solve all of these problems, and serve people really by helping them in searching and downloading software. The system will be a web based expert system for searching and downloading software from all types and usability's.

That expert system will offer classified arrangements for software and powerful searching method through the available software, also information (features, backwards, download instructions, installation instructions… etc), the expert system will be implemented as mentioned above as a web based system, we need to go through several steps to accomplish that goal.

II. RELATED WORKS

In [4] the authors presented the characteristics of disaster management is different in different disaster

relief information, knowledge, standardization of operating procedures and the feasibility of the rescue program. Knowledge-sharing and case retrieval is to develop case-based intelligent decision support system facing the most important issue. Disaster rescue command for decision-making, the use of Web Ontology Language that state the information and knowledge of characteristics of the earthquake disaster rescue, a case-based reasoning and logic to describe the rescue planning business processes.

In [5] presented the main ramification of the idea of "data shouldn't be dumb" on rules languages which is the language has to be simple and able to cover simple linkage situations. More complex applications of rules should be reserved for the intelligent applications that will populate the web. This isn't to say that research and standardization on more powerful and flexible rule languages aren't valuable or shouldn't be done, it's not just necessary or even desirable for the rule-based infrastructure of the semantic web. Keeping these things separate – the infrastructure from the residents of the web – lets us have a much simpler rule language as part of the web infrastructure.

This paper [6] has proposed two extensions for the current generation of digital libraries. First, authors proposed to use ontology to represent scholarly information in digital libraries, thus making the libraries be able to share and exchange knowledge in the Semantic Web environment. Second, fuzzy theory is employed to process uncertain scholarly information as the forms of fuzzy ontology and fuzzy queries. A general architecture of digital libraries in the Semantic Web environment has been presented and an experimental system has also been developed to verify our ideas and techniques.

In [7] proposed an ontology-based information retrieval model to improve effectiveness of information retrieval. The ontology embedded in the proposal model is a fuzzy taxonomy generated automatically from the documents.

In [8] presented an approach to level one sensor fusion in the context of the Semantic Web has been presented using a Semantic Web Expert System Shell (SWEXSYS). Our approach allows for the delegation of sensor fusion tasks to agent-based systems like SWEXSYS. In our approach, we discussed the merit of having Semantic Web enabled data. By making data Semantic Web compliant, an agent may be capable of disambiguating the information; also, additional information that may be needed in the fusion process can be retrieved without human intervention from the Semantic Web by the fusion agent.

III. KNOWLEDGE IN THE EXPERT SYSTEM

Knowledge acquisition, knowledge representation, knowledge storage and knowledge application constitute the main content of expert system's work progress [9], whose core is knowledge base and inference engine. The level of an expert system is basically determined by its knowledge base.

Expert's experience and knowledge are stored in the knowledge base, the more complete and true of the knowledge, the higher level of expert system is [10]. Thus whether it possesses plenty of knowledge is the key of expert system.

There are representation techniques such as frames, rules, tagging, and semantic networks which have originated from theories of human information processing. Since knowledge is used to achieve intelligent behavior, the fundamental goal of knowledge representation is to represent knowledge in a manner as to facilitate inference (i.e. drawing conclusions) from knowledge [11, 12].

We got the following software categories that will be written inside the semantic network knowledge representing as follows:

Internet browser Figure (1), Painting program, Downloading program, Recovery Program Figure (2), Compressors, File sharing Figure (3), Security, Video and audio players, E-books reader, Text Editors, Slides viewers, Images viewers, Cleaning programs

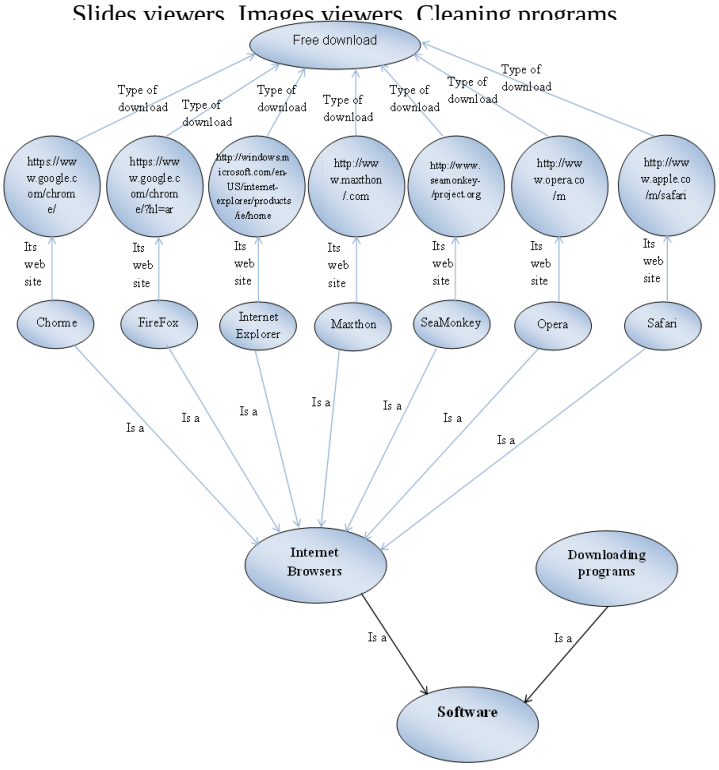


Figure 1: Knowledge Representation for Internet Browser

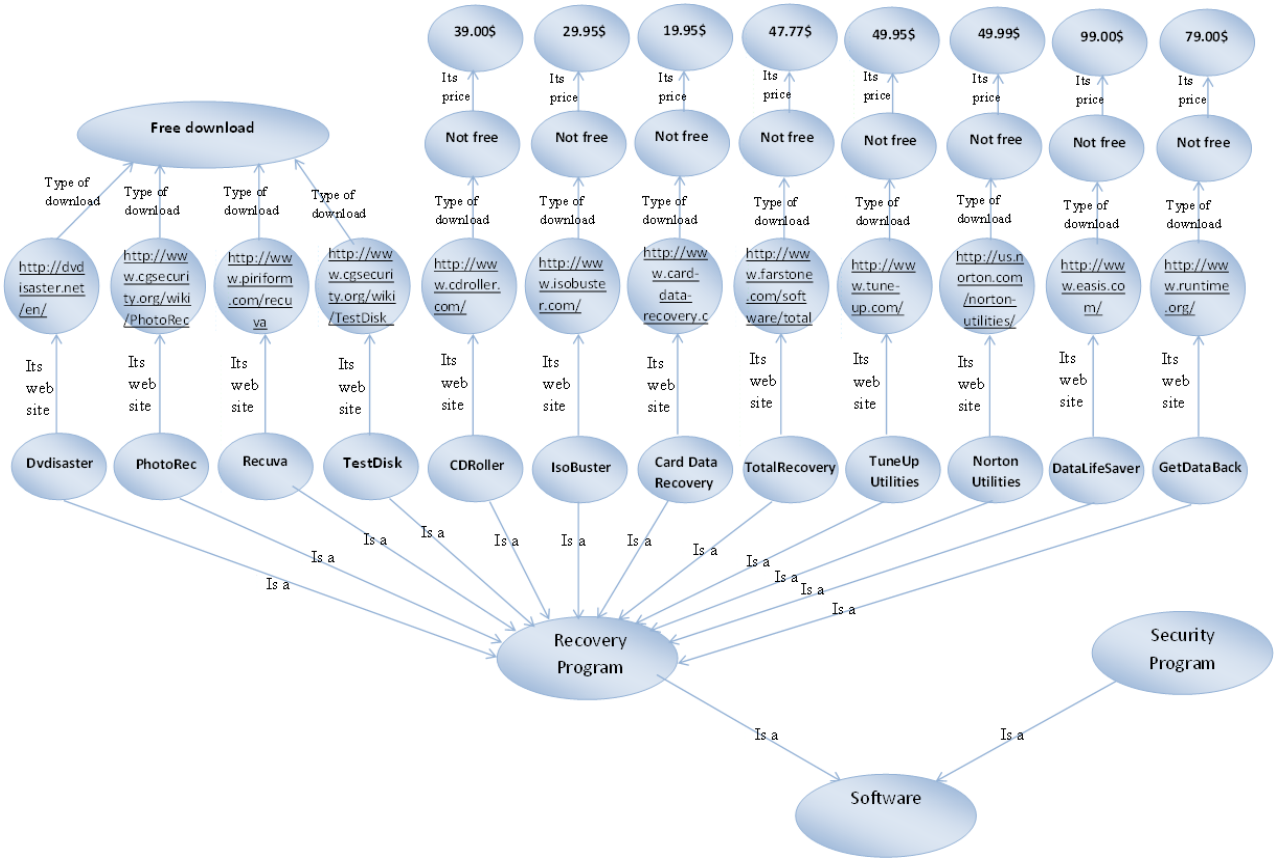


Figure 2: Knowledge Representation for Recovery Program

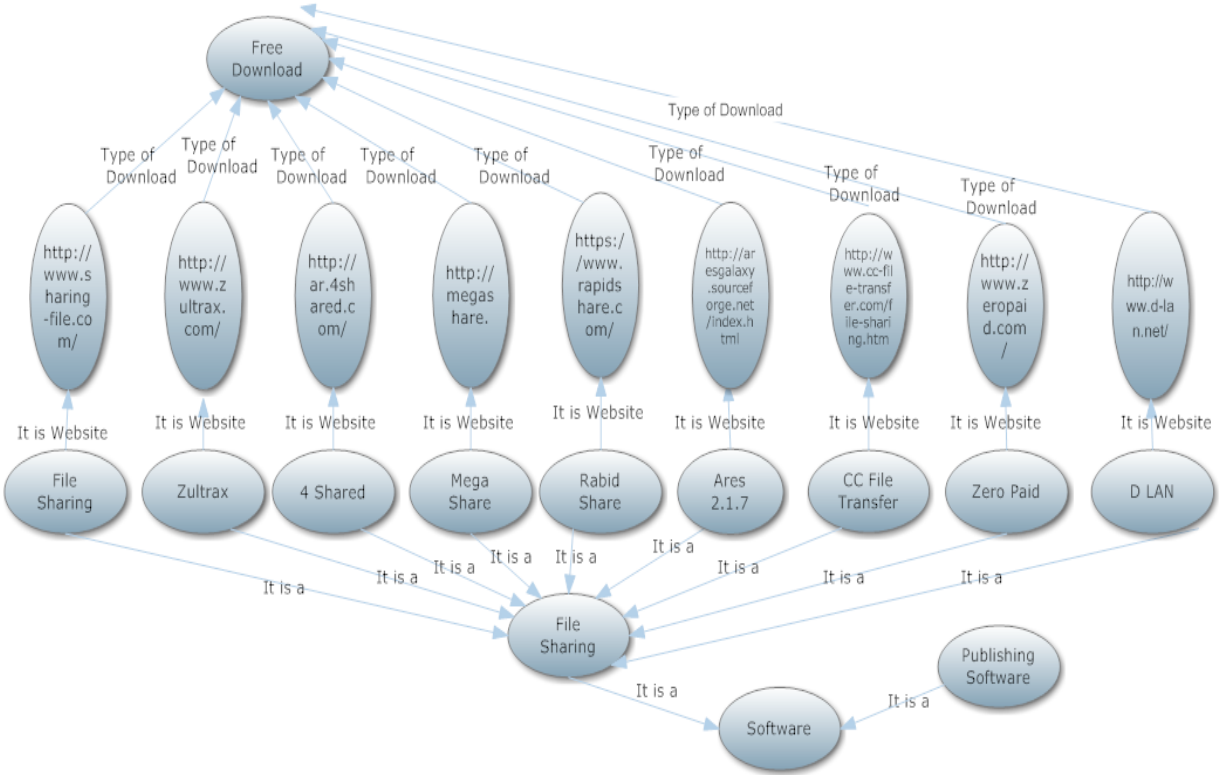


Figure 3: Knowledge Representation for File Sharing Programs

IV. ONTOLOGIES IN SEMANTIC WEB

Ontology is a data model, which can be used to describe a set of concepts and the relationships between those concepts within a domain. Ontology works as the main component in knowledge representation for the Semantic Web. Research groups in both America and Europe developed Ontology modeling languages as The DARPA Agent Markup Language (DAML) and Ontology Inference Layer (OIL.(The W3C Web Ontology Working Group has considered DAML+OIL as the starting point for the introduction of standardized and accepted ontology language for the Semantic Web as Web Ontology Language (OWL) [13]. OWL has three sublanguages: OWL Full, OWL DL and OWL Lite [14]. Existing Semantic Web ontology can be grouped into the following four major categories: metaontology، comprehensive, upper ontology, systematic domain specific ontology, and simple specialized ontology[15].

V. SYSTEM ARCHITECTURE

The system further combines aspects of three research topics (XML, Semantic Web, and Expert System) to facilitate the finding of semantic web services for a client Figure (4).

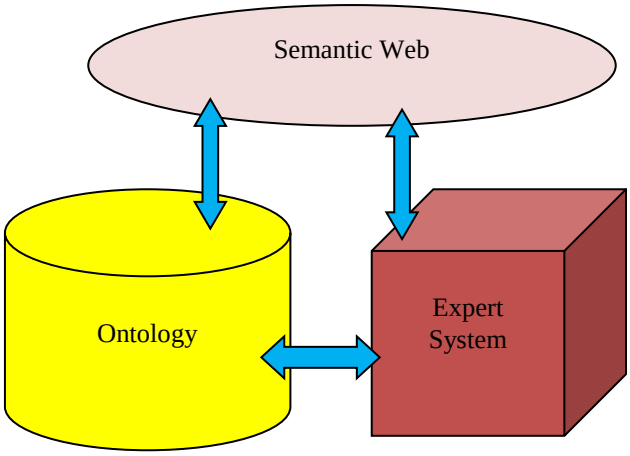


Figure 4: System Architecture

A. XML

Extensible Markup Language (XML) is a set of rules for encoding documents in machine-readable form. It is defined in the XML 1.0 Specifications produced by the W3C, and several other related specifications, all gratis open standards.

The design goals of XML emphasize simplicity, generality, and usability over the Internet. It is a textual data format with strong support via Unicode for the languages of the world. Although the design of XML focuses on documents, it is widely used for the representation of arbitrary data structures, for example in web services.

Many application programming interfaces (APIs) have been developed that software developers use to process XML data, and several schema systems exist to aid in the definition of XML-based languages.

Currently, web services interact by passing XML data, with data types specified using XML Schema. Simple Object Access Protocol [SOAP] can be used as the communication protocol [16], and the I/O signatures for web services are given by Web Services Description Language [WSDL] [17]. UDDI stands for Universal Description, Discovery and Integration [18] and provides the means to publish and discover web services through a UDDI registry.

Today, XML is invading the world of computers and occupying most of its fields. It is widely spreading over the internet, networks, information systems, software and operating systems, DBMS, search tools, web development and services, communication protocols and other fields. As a result, XML data are floating within and between different applications and systems all over the internet and intranets.

A core data representation format for semantic web is Resource Description Framework (RDF). RDF is a framework for representing information about resources in a graph form. It was primarily intended for representing metadata about WWW resources, such as the title, author, and modification date of a Web page, but it can be used for storing any other data. It is based on triples subject-predicate-object that form graph of data. All data in the semantic web use RDF as the primary representation language. The normative syntax for serializing RDF is XML in the RDF/XML form. Formal semantics of RDF is defined as well.

B. Semantic Web Architecture

The architecture of semantic web is illustrated in the Figure (5) below. The first layer, URI and Unicode, follows the important features of the existing WWW. Unicode is a standard of encoding international character sets and it allows that all human languages can be used (written and read) on the web using one standardized form. Uniform Resource Identifier (URI) is a string of a standardized form that allows to uniquely identifying resources (e.g., documents). A subset of URI is Uniform Resource Locator (URL), which contains access mechanism and a (network) location of a document. Another subset of URI is URN that allows identifying a resource without implying its location and means of dereferencing it - an example is urn. The usage of URI is important for a distributed internet system as it provides understandable identification of all resources. An international variant to URI is Internationalized Resource Identifier (IRI) that allows usage of Unicode characters in identifier and for which a mapping to

URI is defined. In the rest of this text, whenever URI is used, IRI can be used as well as a more general concept.

RDFS have semantics defined and this semantics can be used for reasoning within ontology and knowledge bases described using these languages. To provide rules beyond the constructs available from these languages, rule languages are being standardized for the semantic web as well. Two standards are emerging - RIF and SWRL.

For querying RDF data as well as RDFS ontology with knowledge bases, a Simple Protocol and RDF Query Language (SPARQL) are available. SPARQL is SQL-like language, but uses RDF triples and resources for both matching part of the query and for returning results of the query. Since RDFS are built on RDF, SPARQL can be used for querying ontology and knowledge bases directly as well.

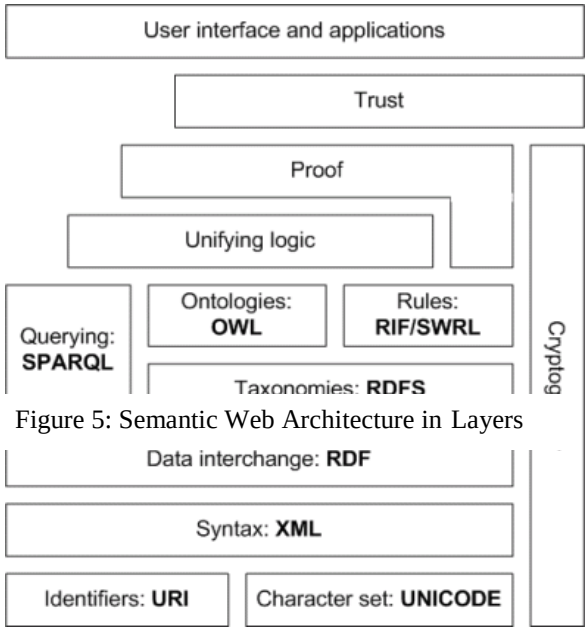


Figure 5: Semantic Web Architecture in Layers

Figure 5: Semantic Web Architecture in Layers

Note that SPARQL is not only query language; it is also a protocol for accessing RDF data. It is expected that all the semantics and rules will be executed at the layers below Proof and the result will be used to prove deductions. Formal proof together with trusted inputs for the proof will mean that the results can be trusted, which is shown in the top layer of the figure above. For reliable inputs, cryptography means are to be used, such as digital signatures for verification of the origin of the sources. On top of these layers, application with user interface can be built.

C. Expert System

An expert system is a computer program designed to simulate the problem-solving behavior of a human who is an expert in a narrow domain or discipline. An expert system is normally composed of a knowledge base (information, heuristics, etc.), inference engine (analyzes the knowledge base), and the end user interface (accepting inputs, generating outputs). The concepts for expert system development come from the subject domain of artificial intelligence (AI), and require a departure from conventional computing practices and programming techniques.

One of the most powerful attributes of expert systems is the ability to explain reasoning. Since the system remembers its logical chain of reasoning, a user may ask for an explanation of a recommendation and the system will display the factors it considered in providing a particular recommendation. This attribute enhances user confidence in the recommendation and acceptance of the expert system.

All expert systems are composed of several basic components: a user interface, a database, a knowledge base, and an inference mechanism. Moreover, expert system development usually proceeds through several phases including problem selection, knowledge acquisition, knowledge representation, programming, testing and evaluation.

1. Inference Engine

The entire control and operation of the system are done by the inference engine; that is developed using C#; which handles the knowledge in format of XML to get the result from the XML file (Knowledge base) that stores the knowledge rules. Figure (6) shows the block diagram of inference engine. The main roles of the inference engine are summarized as: It applies the expert domain knowledge to what is known about the present situation to determine new information about the domain. The inference engine is the mechanism that connects the user inputs in the form of answers to the questions to the rules of knowledge base and further continues the session to come to conclusions. This process leads to the solution of the problem. The inference engine also identifies the rules of the knowledge base used to get decision from the system and also forms the decision tree.

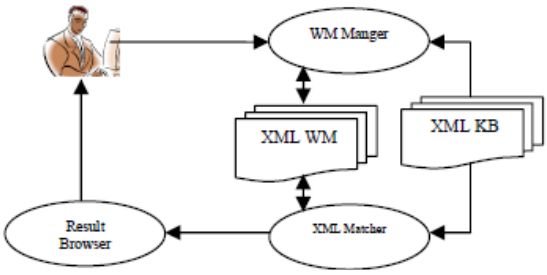


Figure 6: Inference Engine Block Diagram

In this paper the model we took the domain of software is an XML Semantic Web website, we implement the data knowledge based in the form of XML – RDF file to contain all the collected software which we will include in the website, the format of the RDF file as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
- <Data>
  - <SoftwareItem>
      <Category>Internet Browsers</Category>
      <Name>Chrome</Name>
      <Link>https://www.google.com/chrome/</Link>
      <Price>Free</Price>
    </SoftwareItem>
  - <SoftwareItem>
      <Category>Internet Browsers</Category>
      <Name>Fire Fox</Name>
      <Link>http://www.mozilla.org/en-US/firefox/new/</Link>
      <Price>20</Price>
    </SoftwareItem>
  - <SoftwareItem>
      <Category>Internet Browsers</Category>
      <Name>Internet Explorer</Name>
      <Link>http://web-development.techweb.com/?kw=sem_ddj_win_goog_
      <Price>Free</Price>
    </SoftwareItem>
  - <SoftwareItem>
      <Category>Internet Browsers</Category>
      <Name>MaxThon</Name>
      <Link>http://www.maxthon.com/</Link>
      <Price>Free</Price>
    </SoftwareItem>
  - <SoftwareItem>
      <Category>Internet Browsers</Category>
      <Name>Sea Monkey</Name>
      <Link>http://www.seamonkey-project.org/</Link>
      <Price>Free</Price>
    </SoftwareItem>
```

Figure 7: Represented Knowledge in XML.

The knowledge is represented in XML format, because it gives flexibility and being industry standard. Figure (7) shows the part of represented knowledge in XML.

We define 4 attributes:

- Category – Name – Link - Price

VI. RESULTS AND TEST THE SYSTEM

The system was evaluated with different users, including developers, and staff. The system has validated by experts in the domain of online search. Tests of the system were carried out by the developers to make sure the system would work correctly as well as the system is web based system. Figures (8, 9, 10, and 11) shows the snapshots of the developed system.

Figure 8: Home Page

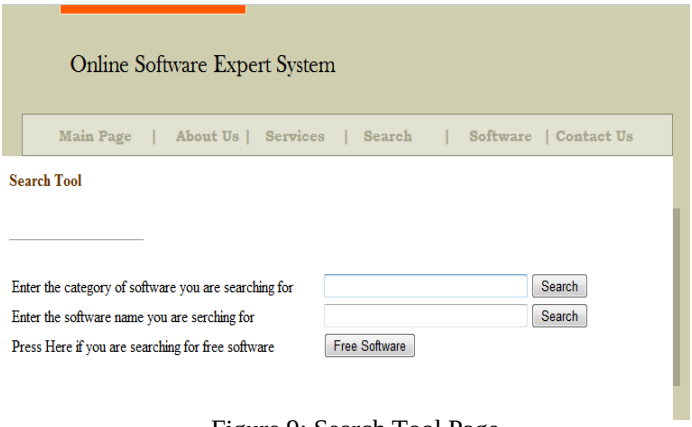


Figure 9: Search Tool Page

Figure 10: All Available Software Page

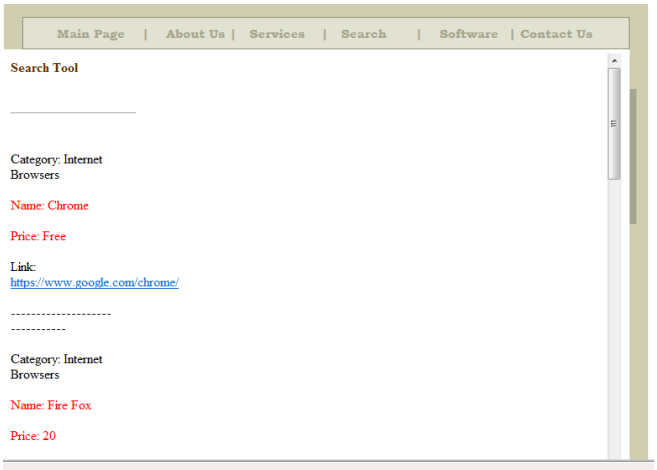


Figure 11: Search Results Page

VII. CONCLUSIONS

In this paper, an expert system based on semantic web for online search has been proposed. It helps users to find the desired software and read about its features and specifications. The expert system saves users' time and effort consumed in web searching. Users can access and react with the proposed system easily using their web browsers. In addition, the proposed system produces interactive on-line objects that provide searching details and advices exactly like to setup a session with human expert. Furthermore, the proposed system combines three main research topics; Semantic Web, Expert System, and XML. Also, organizations can optimize their most valuable expert knowledge through the proposed powerful interactive Web-enabled knowledge automated expert system. Finally, online sessions which have been constructed by the proposed system emulate a human expert conversation, which receives questions to produce customized recommendations.

REFERENCES

[1] Khumukcham R., Shikhar Kr. S, "JESS Based Expert System Architecture for Diagnosis of Rice Plant Diseases: Design and Prototype Development", 2013 4th International Conference on Intelligent Systems, Modeling and Simulation, P 674-676, 2013.

[2] Valerie Cross, Xueheng Hu., "Fuzzy Set and Semantic Similarity in Ontology Alignment", WCCI 2012 IEEE World Congress on Computational Intelligence, Brisbane, Australia June, 10-15, 2012.

[3] Tanimoto, S. L. , " The Elements of Artificial Intelligence " , Computer Science Press. , 1990.

[4] Wu Yun, Li Chan, "Semantic Web-based Seismic Disaster Management Expert System", Project 70601011 supported by National Natural Science Foundation of China, 2009.

[5] Dean Allemang, "Rule-based Intelligence in the Semantic Web ", -or- "I'll settle for a web that's just not so dumb!", Proceedings of the Second International Conference on Rules and Rule Markup Languages for the Semantic Web (RuleML'06), 2006.

[6] Q. Tho, H. Siu, and C. Tru , "Ontology-Based Fuzzy Retrieval for Digital Library", ICADL 2007, LNCS 4822, pp. 95–98. Springer-Verlag Berlin Heidelberg, 2007.

[7] C. Been-chian, H. Chih-hung, J. Ming-yi, "Intelligent Information Retrieval Applying Automatic Constructed Fuzzy Ontology", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007. 1-4244-0973-X/07, IEEE, 2007.

[8] Omoju Thomas, David J. Russomanno, "Applying the Semantic Web Expert System Shell to Sensor Fusion using Dempster-Shafer Theory", 0-7803-8808-9/05, IEEE,2005.

[9] Feng Ding., "Neural Network Expert System", Beijing: Science Press, pp.1-10, 2006.

[10] Wang Hua, Li Peng-bo. "Fault Diagnosis Expert System of Inertial Nacigation System Based on CLIPS and .NET Platform [J]". Journal of Chinese Inertial Technology, Vol.14 No.6, pp.78-80, 2006.

[11] George F. Luger and William, A. Stubblefield; " Artificial Intelligence and the Design of Expert Systems", The Benjamin/ Cummings publishing Co., Inc., 1989.

[12] Keith D., "The Essence of Expert Systems", Prentice Hall, 2000.

[13] G. Antoniou, F. V. Harmelen, "Handbook on Ontologies in Information Systems", Springer-Verlag , pp. 76-92, 2003.

[14] L. Deborah McGuinness, F. V. Harmelen, "OWL Web Ontology Language Overview", W3C Recommendation , Tech. Rep., 10, Feb, 2004.

[15] Li Ding, et al., "Using Ontologies in the Semantic Web: A Survey", Department of Computer Science and Electrical Engineering , University of Maryland Baltimore County, Baltimore, Tech. Rep. TR CS-05-07, July, 2005.

[16] http://www.w3.org/TR/soap/

[17] http://www.w3.org/TR/wsdl.html

[18] http://www.uddi.org/

AUTHORS BIOGRAPHY

DR. Yasser Ahmed Nada Was born in Ismailia, Egypt,

in 1968. He received the BSc degree in pure Mathematics and Computer Sciences in 1989 and MSc degree for his work in computer science in 2003, all from the Faculty of Science, Suez Canal University, Egypt. In 2007, he received his Ph.D. in Computer Science from the Faculty of Science, Suez Canal University, Egypt. From September 2007 until now, he worked as Assistant Professor of computer science. Chair, Department of computer science, Faculty of Computers and Information Technology, Taif University, KSA. His research interests include Expert Systems, Artificial Intelligence, Object Oriented Programming, Computer Vision, and Genetic.

Development of Intelligent Surveillance System Focused on Comprehensive Flow

Shigeki Aoki Tatsuya Gibo† Eri Kuzumoto‡ Takao Miyamoto
Osaka Prefecture University
1–1, Gakuen-cho, Nakaku, Sakai, Osaka, 599–8531 Japan
Email: shigeki_aoki@m.ieice.org
† Hitachi, Ltd, ‡ NEC Corporation

***Abstract*—Surveillance cameras are today a common sight in public spaces and thoroughfares, where they are used to prevent crime and monitor traffic. However, human operators have limited attention spans and may miss anomalies. Here, we develop an intelligent surveillance system on the basis of spatio-temporal information in comprehensive flow of human traffic. The comprehensive flow is extracted from optical flows, and anomalies are identified on the basis of the spatiotemporal distribution. Because our system extracts only a few anomalies from many surveillance cameras, operators will not miss the important scenes. In experiment, we confirmed effectiveness of our intelligent surveillance system.**

***Keywords*—Intelligent surveillance system; comprehensive flow; optical flow; Shannon's information theory**

I. Introduction

Over the recent years, a large number of surveillance cameras have been installed to prevent crimes and to monitor traffic. However, existing surveillance systems rely on a few operators to constantly monitor many cameras. This is impractical as humans can easily miss important scenes. There are two typical routes to solving this problem: One is automated detecting of abnormal behavior[1], [2], [3] and the other is automatic reduction of the number of videos that need to be reviewed[4], [5].

Wu et al. reported that abnormal behaviors can be detected by named Global force and Local force to the trajectories of pedestrians crowded[1]. Gibo et al. proposed a method for recognizing the comprehensive behavior of crowds and detecting anomalies by focusing on optical flows[2]. Mahadevan et al. proposed a method for detecting the anomaly behavior by using a model is based on mixtures of dynamic textures[3]. Alternatively, Saligrama et al. reported a method for analyzing the behaviors of pedestrians and vehicles using global spatial and temporal statistical dependencies of them[4]. Lee et al. proposed a method for recognizing traffic by tracking vehicles across multiple surveillance cameras[5]. However, these methods have several disadvantages. Methods for detecting abnormal behavior[1] can only be applied to certain environments and behaviors that deviate from the SRE model. Moreover, the method used for detecting anomaly behavior [2] cannot detect all abnormal behaviors. Finally, methods that analyze human and vehicular traffic place restrictions on the positions of surveillance cameras[4], [5].

In this paper, we propose a method of detecting anomalies from surveillance video by applying Shannon's information theory to the comprehensive flow of pedestrians[6]. And we report novel experimental results of an intelligent surveillance system on the basis of our method. Since our system detects anomalies from many surveillance cameras, the burden on the human operators is reduced.

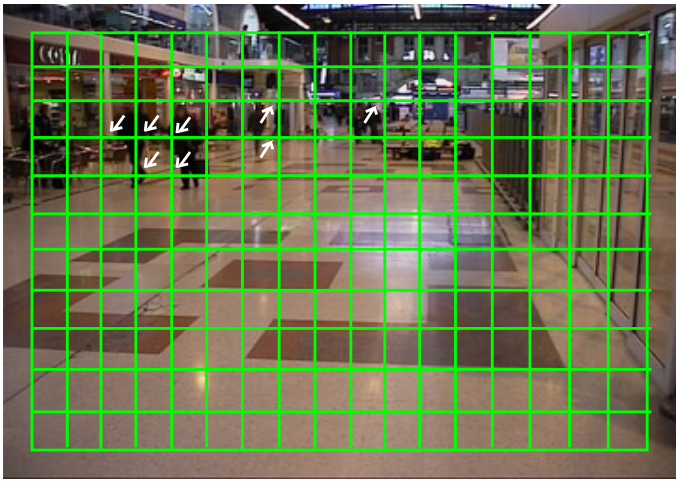


Fig. 1. Mode value of optical flows and image divided into 20×20[pixel] segments. The mode value is drawn in center of each segment. The arrows show the flow directions.

II. Information of Comprehensive Flow

A. Extraction of Comprehensive Flow

We extract the comprehensive flow using the optical flow because it is difficult to keep track of people individually as occlusion occurs frequently in crowded places. That is, we do not observe the individual motions of people, but we observe their comprehensive flow by the following procedure.

Firstly, to extract the comprehensive flow, we calculate the optical flow of each pixel in the input image. Secondly, we divide the image into 20×20 [pixel] segments. This segment size was selected because it can capture the motion of a single person as optical flow. Finally, in each segment, we calculate the mode values of the optical flows. We describe this as the comprehensive flow. Moreover, we exclude the 10-pixel-wide region at the boarder of the input image to eliminate the noises of the optical flows, as shown in Fig. 1.

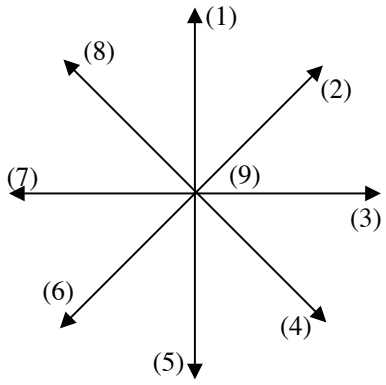


Fig. 2. Directions of quantized comprehensive flows.

B. Probability of Comprehensive Flows

We define anomalies as regions where the flow differs either spatially, that is, with its neighboring regions, or temporally.

Firstly, we quantize the comprehensive flow calculated in Section II-A into nine directions, as shown in Fig. 2. A comprehensive flow with a zero vector is denoted by (9), as shown in Fig. 2. Secondly, we define the appearance probability of a comprehensive flow for a given segment $i(i = 1, 2, ..., N$; where N is the number of regions and $N = 187$.) based on the distribution of comprehensive flows across the neighboring segments. The appearance probability p_i is given by following equation:

$$p_i = \frac{m}{n+1}. \tag{1}$$

where n denotes the number of neighboring segments, m is the number of segments that have the same quantized comprehensive flow as segment i.

C. Spatial and Temporal Information[7]

Here, let us define the information H_i contained by the quantized comprehensive flow of segment i as follows:

$$H_i = -\log_2 p_i. \tag{2}$$

This definition is based on the definition of information in Shannon's information theory. Here, H_i increases as p_i decreases.

We can examine the distribution of the information by calculating the information of each segment in an input image.

In this paper, we discuss two kind of information: spatial variation of the quantized comprehensive flows in several neighboring segments and temporal variation of the quantized comprehensive flows for a fixed segment. The spatial information is given by the variation of the quantized comprehensive flows in each segment of the input image at time t; it generally becomes very large when the flow of people in the image becomes complicated. However, the temporal information is defined by the variation of the quantized comprehensive flows with time for the segment i. We examine the relationship between the spatial and temporal information, as shown in Fig. 3. The figure shows $a \times b$ segments along the spatial direction

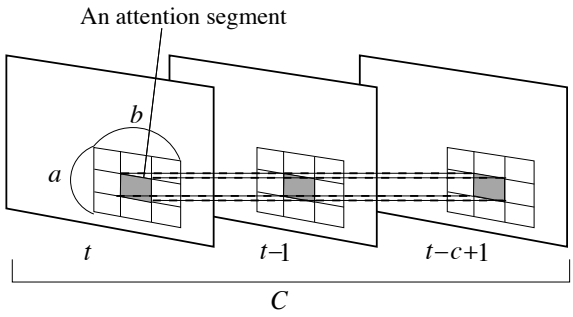


Fig. 3. Neighboring and consecutive region.

Fig. 4. Example of extracted images. The upper left, upper right and lower right images need not be monitored. The lower left image should be monitored.

on one frame and c segments along the temporal direction. For the $a \times b$ neighboring segments on a frame, H_{s_i} represents the spatial information. And for the c segments along the time axis at i, the temporal information is given by H_{t_i}.

D. Combined Spatial and Temporal Information[7]

We consider the combined spatial and temporal information I_i to detect anomalies. I_i is given by the following equation:

$$I_i = H_{s_i} + H_{t_i} - H_{st_i}. \tag{3}$$

where H_{st_i} denotes the spatiotemporal information of the segment i in the domain of the neighboring $a \times b \times c$ segments, where $a \times b$ segments are in the spatial direction and c is the number of consecutive frames, as shown in Fig. 3.

III. Detection of Anomalies

Firstly, we calculate the mutual information I_i from each segment i for each surveillance camera. Secondly, we calculate the sum of mutual information as follows:

$$\mathcal{I} = \sum_{i=1}^{N} I_i. \tag{4}$$

Fig. 5. Example of images of each dataset.

Fig. 6. Example of experimental result. This result denotes correct extract of anomalous image and anomalous region. The upper right image should be monitored.

In above equation, when $\mathcal{I}$ is large, the input image must be reviewed by humans. To determine this, we define the thresholds α and β. When $\mathcal{I} < \alpha$, we draw a white circle on the top of the image that indicates that the image need not be monitored. When $\alpha \leq \mathcal{I} < \beta$, we draw a gray circle, and when $\mathcal{I} \geq \beta$, we draw a black circle. The black circle indicates that the image should be monitored. Footage for which a black circle is presented anomalous image. Moreover, we define segments where I_i is larger than a threshold γ as anomalous region. Such regions are highlighted in the video.

As shown in the example in Fig. 4, rectangle highlight anomalous region and the color of the circle identifies anomalous image. In Fig. 4, the upper left, upper right and lower right images need not be monitored. Only the lower left image should be monitored. Therefore, our method can extract and display anomalies.

IV. Experiment and Discussions

A. Experimental Condition

We developed an intelligent surveillance system on the basis of the proposed method. In order to verify the effectiveness of the our system, we performed an experiment by using PETS 2006 sample sequences (Scene7-1 and Scene7-2)[8], UCSD anomaly detection dataset[9] and Mall dataset[10]. Fig. 5 shows images of each dataset. Image sizes were reduced to 360×288 pixels to reduce the processing time and filter out the noise. We evaluated the false positive (FP) and the false negative (FN) for 1,950 images available of each dataset. FP means that a video that need not be monitored is incorrectly identified. FN means that a video that must be monitored is not identified. We set the thresholds as follows on the basis of a preliminary experiment: $a = 5$, $b = 5$, $c = 50$, $\alpha = 40$, $\beta = 60$, and $\gamma = 2$.

B. Experimental Results

In Fig. 6, only upper right image must be monitored. This example confirmed that anomalous image could be correctly extracted. Images with crowding were extracted especially well. In addition, we confirmed that crowded regions could be extracted as an anomalous region in Fig. 6.

Fig. 7. Example of noise in the Mall dataset. This result denotes incorrect extract of anomalous region.

First, we evaluated the FP and FN of anomalous images. Table I shows the experimental result of accuracy of anomalous images. We had 62 FP and 68 FN among 7,800 ($1{,}950 \times 4$) images. From the experimental results, we verified that 48 of the 52 FP in Mall dataset were due to noise of the optical flows by tree. Fig. 7 shows an example of noise in Mall dataset.

Second, we evaluate the FP and FN of anomalous regions by visual judgment. Table II shows the experimental result of accuracy of anomalous regions. We had 316 FP and 167 FN among 1,458,600 ($7{,}800 \times 187$) segments. Fig. 8 shows an example of accuracy of anomalous regions. Almost all FP and

TABLE I. EXPERIMENTAL RESULTS OF ACCURACY OF ANOMALOUS IMAGES.

	FP	FN	Total
PETS2006 S7-1	0	1	1 (0.0%)
PETS2006 S7-2	1	2	3 (0.2%)
Mall Dataset	52	61	113(5.8%)
UCSD Dataset	9	4	13 (0.6%)
Total	62 (0.8%)	68 (0.9%)	130 (1.7%)

TABLE II. EXPERIMENTAL RESULTS OF ACCURACY OF ANOMALOUS REGIONS.

	FP	FN	Total
PETS2006 S7-1	15	1	16 (0.0%)
PETS2006 S7-2	5	10	15 (0.0%)
Mall Dataset	195	121	316 (0.1%)
UCSD Dataset	101	35	136 (0.1%)
Total	316 (0.0%)	167 (0.0%)	483 (0.1%)

FN were due to noise of the optical flow.

Thus, this system makes it possible to identify images and regions that should be monitored, which reduces the burden on human operators. We confirmed the effectiveness of our method. However, there are some problems, especially with regard to FP, which are attributable to noise in the optical flows.

C. Discussions

We extract comprehensive flows using optical flows. It is difficult to keep track of people individually because occlusion occurs frequently in crowded places. We confirmed correctly extract comprehensive flows from experimental results. But we could not extract comprehensive flows correctly if the segment size was not suitable for the size of a person in the video. Our future goal is to determine a method to set segment size according to the surveillance environment.

We focus on the complexity of the comprehensive flows. We considered the spatial and temporal information of comprehensive flow. We confirmed that mutual information I_i is large when flows of segment i and neighboring segments are complex.

In order to extract anomalous images, we define the threshold α and β. When $\mathcal{I} \geq \beta$, we draw black circle and indicate that the image should be monitored. Moreover, we define segments where I_i is larger than threshold γ as anomalous region. Such regions are highlighted in the image. Our system demonstrated a 98.3% success rate for extracting anomalous images, and a 99.9% success rate for extracting anomalous regions. Since our system detects anomalies from many surveillance cameras correctly, the burden on the human operators is reduced.

We performed an experiment to provide performance measurements among our method and Mahadevan et al. method[3] using same UCSD anomaly detection dataset[9] which we used for experiments of detection in the above sections. When their method demonstrated an about 50% total of FP and FN rate. In our method, we demonstrated a 0.6% total of the FP and the FN rate.

Fig. 8. Example of experimental result of accuracy of anomalous regions.

V. CONCLUSION

We proposed a method for extracting anomalous images and anomalous regions in surveillance video using Shannon's information theory and the comprehensive flow of pedestrian traffic. And we developed an intelligent surveillance system on the basis of our method. We also confirmed the effectiveness of our intelligent surveillance system. Since our system detects anomalies from many surveillance cameras, the burden on the human operators is reduced. Our future goal is to determine methods to set segment size according to the surveillance environment.

ACKNOWLEGEMENT

This work was supported by JSPS KAKENHI Grant Number 24700195.

REFERENCES

[1] S. Wu, B. E. Moore, M. Shah, "Chaotic Invariants of Lagrangian Particle Trajectories for Anomaly Detection in Crowded Scenes", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1975–1981, 2010.

[2] T. Gibo, S. Aoki, T. Miyamoto, M. Iwata and A. Shiozaki, "Sequential Learning and Recognition of Comprehensive Behavioral Patterns Focused on Confluence", Proceeding of Japan-Cambodia Joint Symposium on Information Systems and Communication Technology (JCAICT), pp. 83-88, Jan. 2011.

[3] V. Mahadevan, W. Li, V. Bhalodia, N, Vasconcelos, "Anomaly Detection in Crowded Scenes", IEEE Proceedings of Computer Vision and Pattern Recognition (CVPR), pp. 1975–1981, 2010.

[4] V. Saligrama, Z. Chen, "Video Anomaly Detection Based on Local Statistical Aggregates", IEEE Proceeding of the Computer Vision and Pattern Recognition (CVPR), pp. 2112–2119, 2012.

[5] L. Lee, R. Romano, and G. Stein, "Monitoring Activities from Multiple Video Streams: Establishing a Common Coordinate Frame, " IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI), Vol. 22(8), pp. 758–767, 2000.

[6] T. Gibo, E. Kuzumoto, S. Aoki, T. Miyamoto, and M. Yoshioka, "Automatic Extraction of Videos of Interest and Regions of Interest from Images of Surveillance", Proc. of The First Asian Conference on Information Systems (ACIS), pp. 29–32, Dec. 2012.

[7] M. Onishi, M. Izumi, K. Fukunaga, "Production of Video Images by Computer Controlled Camera Operation Based on Distribution of Spatiotemporal Mutual Information", Proceeding on 15th International Conference on Pattern Recognition (ICPR), Vol. 4, pp. 102–105, Sep. 2000.

[8] "Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance", ftp://ftp.cs.rdg.ac.uk/pub/PETS2006/S7-T7-A.zip.

[9] "UCSD Anomaly Detection Dataset," http://www.svcl.ucsd.edu/projects/anomaly/dataset.html

[10] "Mall Dataset," http://www.eecs.qmul.ac.uk/~ccloy/files/datasets/mall_dataset.zip

FPGA Architecture for Kriging Image Interpolation

Maciej Wielgosz, Mauritz Panggabean and Leif Arne Rønningen
Department of Telematics
Norwegian University of Science and Technology (NTNU)
N-7491, Trondheim, Norway

***Abstract*—This paper proposes an ultrafast scalable embedded image compression scheme based on discrete cosine transform. It is designed for general network architecture that guarantees maximum end-to-end delay (EED), in particular the Distributed Multimedia Plays (DMP) architecture. DMP is designed to enable people to perform delay-sensitive real-time collaboration from remote places via their own collaboration space (CS). It requires much lower EED to achieve good synchronization than that in existing teleconference systems. A DMP node can drop packets from networked CSs intelligently to guarantee its local delay and degrade visual quality gracefully. The transmitter classifies visual information in an input image into priority ranks. Included in the bitstream as side information, the ranks enable intelligent packet dropping. The receiver reconstructs the image from the remaining packets. Four priority ranks for dropping are provided. Our promising results reveal that, with the proposed compression technique, maximum EED can be guaranteed with graceful degradation of image quality. The given parallel designs for its hardware implementation in FPGA shows its technical feasibility as a module in the DMP architecture.**

I. Introduction

Greater interest in green technology and rapid technological advances opens ways to conceive multi-party collaboration from distributed places via tele-immersive environment. Near-natural quality is achievable by tiling auto-stereoscopic multi-view 3D displays and high-end cameras on all the surfaces of such environment. The collaborations will soon include those which are very sensitive to end-to-end delay (EED), such as remote choir-conducting and dancing. The effect of EED is very critical to achieve good synchronization between the collaborating people from different locations. For example, the optimal EED for synchronizing rhythmic clapping hands from different places is 11.5ms [3]. It includes propagation, transmission and all processing delays. Longer EED will produce increasingly severe tempo deceleration while shorter ones yield a modest yet surprising acceleration. Percussion is rhythmically similar to clapping hands. Musicians playing percussion will then require similar EED for both audio and video data while collaborating. The same effect of EED to collaborative dancing is indicated in [22]. Video data is essential to good synchronization between dancers as they depend on visual cues [3]. Thus the greatest demand to meet and guarantee such low EED lies in processing visual information.

The Internet today, however, is unable to deliver such guarantee with its best-effort design. One approach for it is to design network nodes with ability to intelligently drop video packets on-the-fly despite changing traffic conditions but also with graceful quality degradation. Video compression by current coding standards is conducted only by the sender. Higher visual quality comes at the expense of more complexity and longer encoding time. Therefore very low EED implies minimizing or even avoiding video coding at the expense of high increase in bit rate. It implies the use of intraframe, object-based and parallel processing. To guarantee both the constant EED and the graceful degradation of visual quality, a novel network architecture namely Distributed Multimedia Plays (DMP) has been proposed [16].

Very demanding situations during collaboration, for example when the input video from the collaboration environment is extremely transient, require simplified approach in processing the data. A simple data representation for parallel image transmission is shown in Fig. 1. $N \times N$ blocks are tiled directly on the pixels of a video frame, yielding N^2 bit streams of pixel values. In DMP objects in an image can be segmented, processed and transmitted independently. The number of pixel streams in a segmented object may vary depending on its visual content. Thus network nodes can drop pixel streams after entropy coding to instantly reduce the bit rate according to immediate traffic conditions.

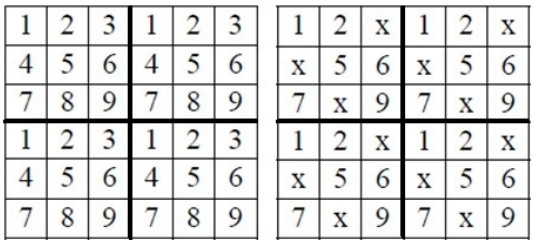


Fig. 1. The tiling of 3×3 blocks ($N = 3$) over an image of 9×9 pixels, yielding $N^2 = 9$ streams of pixels (left); the dropping of streams number 3, 4 and 8 (right). Each pixel of dropped streams denoted by $\times$ will be optimally interpolated from the remaining ones at the receiver.

The time for data reduction by network nodes must also be minimized. This affects how the entropy coding must be designed later. The dropped pixels will be estimated by applying optimal interpolation in the sense of mean square error to the received bit streams at the receiver. Searching for such interpolation leads us to Kriging, a technique well-known and widely used in geostatistics. Kriging works better by using window mechanism, hence called windowed kriging interpolation. It was proposed and used to interpolate luma and chroma data in natural images with positive results [12][13].

The collaboration system on the DMP algorithm must work fast due to the very low EED. Practically it means that the processing routines are to be implemented in hardware. In this work we focus on field-programmable gate array

(FPGA) implementation. To the best of our knowledge, there is no reliable reported information on the feasibility of FPGA architecture for Kriging image interpolation. This work is motivated to fill that void. Some of the calculation routines in the proposed architecture are reused. Moreover some highly exhaustive computations are skipped whenever possible. The paper is structured as follows. Section II presents an in-depth analysis of the Kriging algorithm that leads to the proposed FPGA architecture. The computational complexity and the resource consumption of the architecture are discussed in Section III. Section IV finally concludes the paper with the summary and some ideas for future work.

II. Analysis of the Kriging Algorithm and the Proposed Architecture

The computational cost of some parts of the Kriging algorithm is $O(n^4)$ where n is the size in pixels of the square image interpolated. The strong data dependencies in the algorithm make parallelization a challenge. Some of the inherently sequential stages in the algorithm are best performed on general purpose processor (GPP) rather than on FPGA.

Implementations of the Kriging algorithm on different computational platforms of GPP or graphics processing unit (GPU) have been reported such as in [6][4][8][17]. The speedup factors are promising and span from eight [6][4], twelve [17] and up to 120 [8]. The latter is achieved on GPU by means of compute unified device architecture (CUDA). The architecture for OK presented in this paper aims to decrease the overall latency in FPGAs. Therefore direct use of the aforementioned speedup benchmarks is not adequate for DMP. The goal of most of these implementations is to reach high data-processing throughput rather than low system latency. This is understandable as most of the papers focus on geology and geostatistics [6][4][8][17] instead of image processing. Nevertheless there are some work that cover the latter in [10][5][14] although they do not address FPGA implementation.

Ordinary kriging (OK) is employed in this work to interpolate the dropped pixels in an output image. ML605 platform from Xilinx [18] is chosen for system deployment and estimates. The analysis consists of three computational steps of the OK algorithm [7]. Step 1 is to find points in an image which contribute to building an output picture. Step 2 is to construct the variogram matrix. The last step is to compute the weights and interpolate the missing points.

A. Step 1: Finding Points as Basis for Interpolation

The step for the windowed Kriging interpolation [11] is not very difficult since all the input points to be considered are included in the well-established frame of interest of $D \times D$ pixels. The optimal size of D is a subject of research and will be provided later in this paper. A DMP access-node reconstructs an image out of the received packets according to the employed dropping scheme [1]. The number of dropped packets varies depending on the network traffic. Image data in pixels are gradually delivered to the Kriging interpolation module (KIM) which is responsible for the image reconstruction. An overview of the process is depicted in Fig. 2.

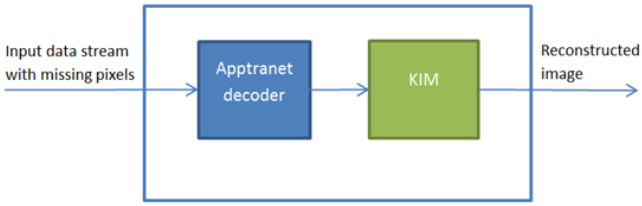


Fig. 2. The image reconstruction scheme in DMP.

B. Step 2: Constructing the Variogram Matrix

A variogram provides the information on the contribution of a given point to that being reconstructed [7]. It is a function of a distance vector between the points $\mathbf{h}$ as $\gamma(\mathbf{h}) = \frac{1}{2}E[(z(\mathbf{s}_i+\mathbf{h})-z(\mathbf{s}_i))^2]$ where $z(s_i)$ are observations in two different locations separated by $\mathbf{h}$. A variogram is obtained from an experimental semivariance as a result of its fitting to one of variogram models [7], [11]. Experimental semivariance for each distance $\mathbf{h}$ in an interpolated image frame may be obtained as

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} E[(z(\mathbf{s}_i + \mathbf{h}) - z(\mathbf{s}_i))^2] \qquad (1)$$

where $N(\mathbf{h})$ denotes the number of sample pairs separated by $\mathbf{h}$ [7]. There are five stages to compute the variogram:

1) Determining the squares of the distances between all the points in the area of interest constituted by the $D \times D$ window
2) Sorting the distances
3) Accumulating the sorted values
4) Building an experimental variogram
5) Fitting the variogram to an appropriate model.

The computational complexity of the first stage can be shown as $D^2(D^2-1)/2 \approx O(D^4)$. The operations in this stage involve both subtraction and multiplication of each pair of the $D \times D$ pixels. Multipliers and adders in Stage 1 are connected in a pipeline fashion to the other modules, as illustrated in Fig. 3. The overall number of different inter-pixel distances within a picture, $N(D)$, is given by

$$N(D) = \binom{D+12}{} + (D-1) = \frac{(2+D-2)!}{2(D-2)!}$$
$$= \frac{D(D-1)}{2} + (D-1) = \frac{(D-1)(D+2)}{2}$$

where D is the length of an image frame edge in pixels. $N(D)$ is plotted against D in Fig. 4.

A separate hardware module can be dedicated to each processing stream, as encircled in black in Fig. 3, leading to $N(D)$ streams. The number of pixels processed by a single stream in such a scheme is expressed by

$$k = \frac{\text{total number of pixels in a picture}}{\text{overall number of distances between pixels}}$$
$$= \left\lceil \frac{D^2(D^2-1)/2}{(D-1)(D+2)/2} \right\rceil$$
$$= \left\lceil \frac{D^2(D+1)}{D+2} \right\rceil \approx D^2.$$

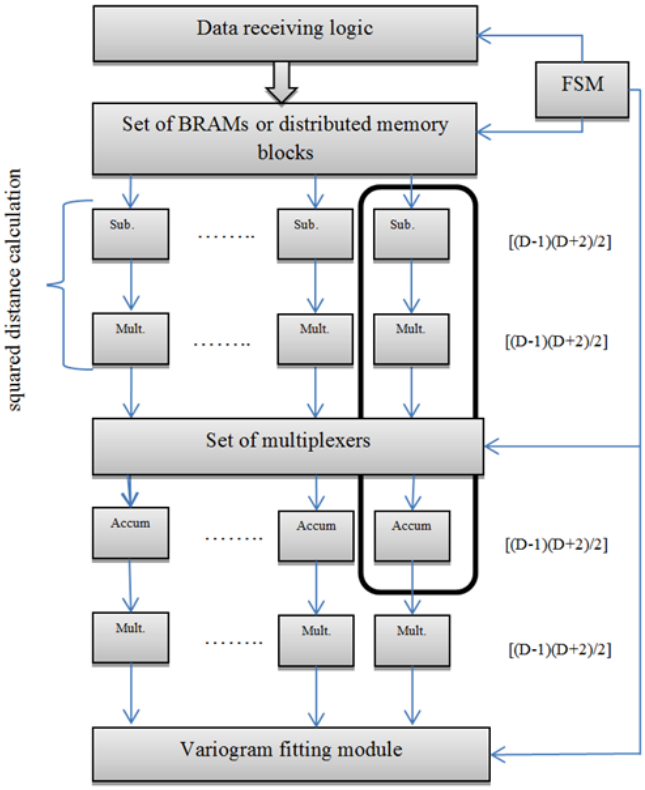


Fig. 3. The proposed architecture of the variogram calculation module.

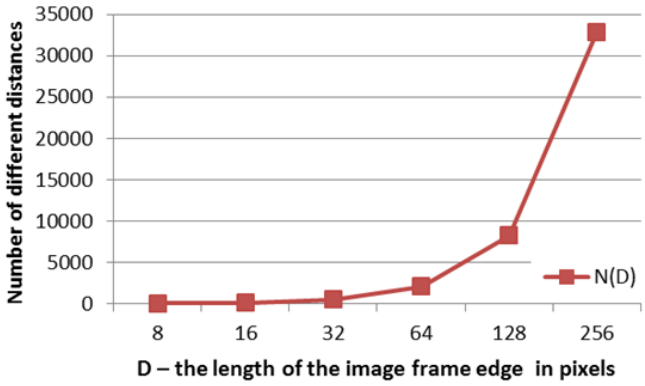


Fig. 4. The number of different distances as a function of D.

Note that this approach leads to large memory consumption since $N(D)$ of k KB Block RAMs (BRAMs) or distributed RAM memory blocks are required. It is because a separate memory port is needed for data fetching and reading. Sorting can be performed during the data fetching or by means of a set of $(N(D)$ to 1) multiplexers (MUXs). The pipeline architecture of the variogram calculation unit presented in Fig. 3 features all the processing stages of Step 2.

There are $4[(D-1)(D+2)/2]$ units in a variogram-calculation module with an approximately 10 clock cycles (CLKs) + MUX pipeline latency. Here the variogram-fitting module is not taken into account. It is assumed that a multiplier consumes 3 CLK while an adder and an accumulator take 1 CLK. The pipeline delay also depends on the data width. To achieve the final precision of 8 bits, some guard bits are used in the middle processing stages (see Fig. 3). However to simplify the analysis, this parameter is set to 0, i.e. no guard bits.

Unfortunately the pixel pairs are not evenly distributed across the streams. For example, from $2D(D-1)$ pairs of neighboring pixels, two are of the maximum distance, i.e. the diagonal of an image frame. Therefore the architecture (without MUX) illustrated in Fig. 3 is not optimal in terms of data distribution. Some of the streams will finish their tasks earlier and remain idle while the others are still being processed. Load balancing procedure may be implemented as finite state machine (FSM) to provide equal distribution of pixels across the streams. This procedure would require driving MUXs that feed the accumulators with data according to the computed semivariance. However, the delay introduced by the MUXs is large, due to their sizes and resource consumption.

An interesting alternative would be time-multiplexing of variogram computations coupled with an accumulator-result summation. Instead of implementing a set of MUXs, Eq. (1) may be multiplexed in time, reusing the streams in Fig. 3 multiple times during computations. The decision regarding the number of parallel streams should consider the expected size of the input image. Fig. 5 plots the variogram computation time for one stream against D. The estimate assumes that the FPGA is clocked at 100 MHz. The remaining part of the paper also adopts that assumption. As mentioned above, computation time depends on the degree of parallelism. In the case of variogram calculation stage, it is strictly related to the number of processing streams employed. Fig. 6 exhibits this relationship, as a function of the number of concurrently working module for D that equals 256, 128 and 64 pixels.

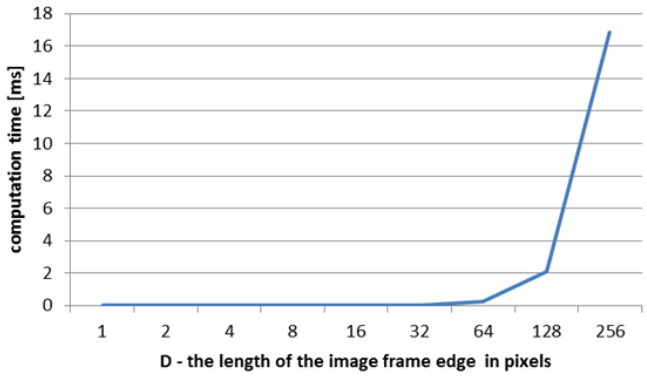


Fig. 5. Computation time of a single stream against D.

The throughput of the external data bus is a key factor in estimating the system performance. There are two communication buses employed in DMP: PCIe and 10Gb Ethernet. The transfer speed of the latter is 2 or 1.25 GB/s. PCIe bus is used for local data transfer between the camera devices and the processing boards. The 10Gb Ethernet provides the inter-node communication [15]. Fig. 7 reflects the expected throughput of the two interfaces for different values of D. Raw transfer that includes the overhead of the protocols is considered here for simplicity. The worst case in Fig. 7 occurs when only one pixel is missing and must be interpolated from the input data. In real situation it will not occur since the smallest amount of data being dropped is $1/N^2$ of all the image pixels. The

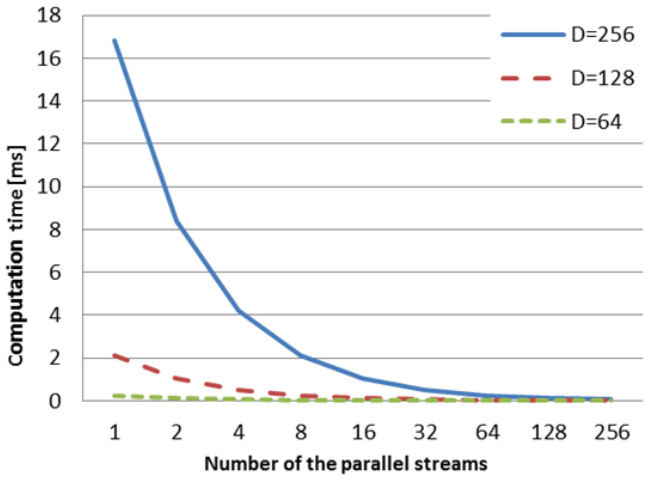


Fig. 6. The impact of the number of concurrently employed streams on the computation time.

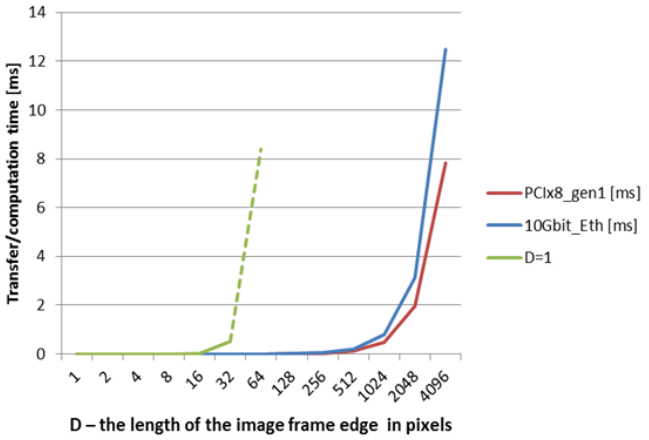


Fig. 7. Data transfer/computational time against D.

number of dropped pixels varies from $1/N^2$ to $(N^2-1)/N^2$, either fully or partially (cf. Fig. 1). Fig. 7 also depicts how the growth of an image size affects both the data transfer and computation time as shown by the dotted green line. One can see that the computation time, rather than the data transfer, is the constraint.

Consequently it is possible to generate a variogram for quite large images given enough parallel streams. Here we focus on 256×256 images. A single stream absorbs roughly 270 look-up tables (LUTs), 252 flip-flops (FFs) and one BRAM [19]. Thus a single Virtex-6 XC6VLX240T chip can accommodate 150,720/270 $\approx$ 558 streams if the time-multiplexing is chosen, i.e. no resource is reserved for implementing MUXs. Accordingly, a module comprising 256 streams capable of handling computations of a 256×256 image consumes roughly 46% of all the FPGA resources. It generates results in approximately 10 CLK or 0.1 μs.

The stream switching tasks in the time multiplexing takes additional time, but completing the computations under 1μs is feasible. It is possible to reduce resource utilization by reducing the number of parallel streams at the expense of higher multiplexing effort. For instance, the consumption of resources at 64 streams drops to approximately 11% and the computation time remains below 1ms, as shown in Fig. 6.

The analysis presented in this section does not cover variogram fitting procedure due to its sequential nature. Moreover the other blocks in Fig. 3 absorb many resources. Thus the computation of variogram matrix is assumed to be made offline and uploaded to the FPGA before the computations start. This section is intended to show that implementing Step 2 on FPGA is feasible. It may be attempted later with high probability. It would be especially beneficial if some of the other modules presented later in the paper, e.g. the linear solver, could be reused for this purpose. It is saved for future research, along with studying the impact of using fixed variogam on the quality of the image interpolation. By fixed variograms we mean using the same variogram matrix for different images, instead of generating a different one for each new incoming frame.

C. Step 3: Computing the Coefficients and Interpolation

This step is performed directly after Step 2 and implemented as a separate hardware module. It is again useful to skip the experimental variogram fitting. Therefore it is assumed here that the variogram matrix is computed offline and uploaded to FPGA internal memory. This does not exclude the possibility of incorporating the generation of the variogram matrix into the system later.

Step 3 features two operations that are implemented as separate hardware modules. The first is the construction of the semivariogram matrix. The second is to solve the linear equation to compute the λ coefficients.

A variogram matrix can be built by gradually filling it up with data derived from the model equations. The following exponential model is the most suitable for image interpolation

$$f(h) = c\left(1 - e^{-3h/a}\right) \tag{2}$$

where h, a, c are the parameters computed in Step 2 or uploaded to FPGA memory [16]. The model can be implemented as hardware module that consists of a *exp()* unit, a subtractor and multipliers, as pictured in Fig. 8.

Parameters h, a, c and $ln(c)$ are computed offline, e.g. on a GPP. Since there are only four parameters, it is possible to store a few sets of them in the internal memory and use a pointer to the selected one. The selection could be made based on the properties of each incoming image. To reduce resource consumption, the last multiplier in Fig. 8 (b) can be eliminated by computing $ln(c)$ as shown in Fig. 8 (a).

If the variogram matrix is implemented as a LUT memory, the number of the entries equals $N(D)$ as shown in Fig. 4. Thus it is important to take into account the available memory resources. For example, 256×256 pixel image occupies roughly 1% of the internal BRAM memory resources of the Xilinx Virtex-6 XC6VLX240T. Depending on the available FPGA resources, one can adopt a strategy by either calculating the variogram values using Eq. (2) or storing a computed variogram matrix in the internal memory. However, regardless of how the variogram matrix is generated, it is passed on to the linear solver.

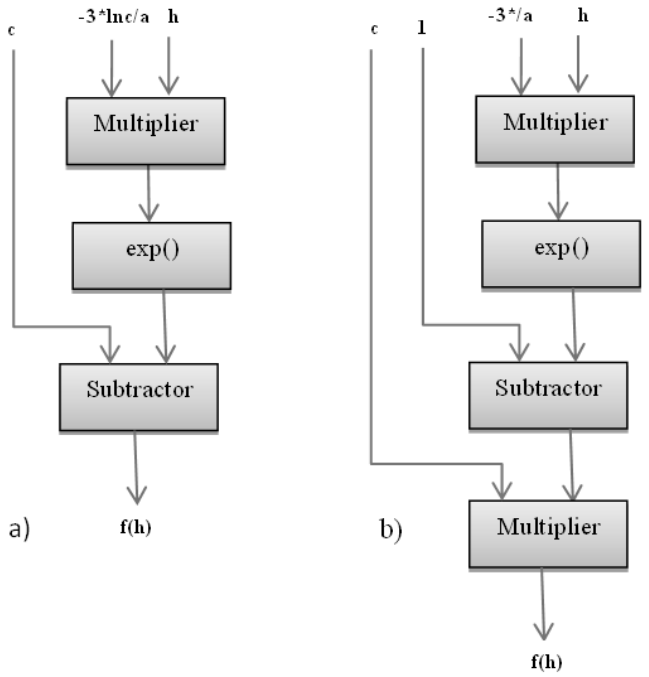


Fig. 8. Two block diagrams of the $f(h)$ module.

1) The linear equation solver: The goal of the Kriging procedure is to interpolate the missing image pixels at the DMP receiver. Thus, based on the variogram matrix generated in the previous step of the algorithm, it is essential to compute the λ_i coefficients for every missing point p separately as

$$p = \sum_{i=1}^{n} \lambda_i \mathbf{w}(x_i) \tag{3}$$

where λ_i and $\mathbf{w}(x_i)$ are the coefficient vectors and the values at the known points, respectively [11]. The equation for computing λ_i in matrix formulation is given by

$$\begin{bmatrix} \lambda_i & 1 \\ 1^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} \lambda_p \\ 1 \end{bmatrix} \tag{4}$$

where λ_i is the variogram matrix and λ_p is the semivariance vector of the points being interpolated. Since Eq. (4) is solved for all the missing points using the same variogram matrix, one can compute the inverse matrix or use one of the well-known factorization schemes to simplify the calculations, such as QR, LU or Cholesky decompositions. Although the computational complexity of these transformations is $O(n^3)$, the cost in hardware implementation is different. Neither QR nor LU takes the symmetry of the matrix directly into account. It leads us to use Cholesky decomposition. FPGA implementation of Cholesky decomposition has been studied and reported in many papers such as [9][20][2].

2) Cholesky decomposition: Cholesky factorization can be performed for symmetric positive definite matrices and this is the case for the variogram matrix. The decomposition scheme is given by

$$A = L\,L^T \tag{5}$$

where A and L are the input matrix and the lower triangular matrix, respectively. It is worth noting that only one matrix L is to be stored in the memory and used. The data throughput of the module can be efficiently doubled. The example of Cholesky transformation for a 3×3 matrix is as follows:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{12} & l_{13} \\ 0 & l_{22} & l_{23} \\ 0 & 0 & l_{33} \end{bmatrix}.$$

Multiplying the matrices in Eq. (5) and solving the set of six equations by forward-substitution lead to the formula that defines an algorithm for a general case as

$$a_{jk} = \sum_{s=1}^{k} l_{js} \cdot l_{ks} \quad \text{for } 1 \le k \le j \le n \tag{6}$$

where j, k, s are matrix indices. Eq. (6) can be defined as

$$l_{jk} = \frac{1}{a_{kk}} \left(a_{jk} - \sum_{s=1}^{k-1} l_{js} \cdot l_{ks} \right) \quad \text{for } j > k$$

$$l_{kk} = \sqrt{a_{kk} - \sum_{s=1}^{k-1} l_{ks}^2} \tag{7}$$

The square root occurrence in Eq. (7) is a drawback from hardware implementation perspective. The computation order exhibits data inter-dependencies as shown by

$$\begin{bmatrix} 1 & & & \\ 2 & 3 & & \\ . & . & 5 & \\ 2 & 4 & ... & 2n-1 \end{bmatrix}.$$

The disadvantages can be overcome by incorporating a third matrix D as

$$A = L\,D\,L^T. \tag{8}$$

The example of 3×3 Cholesky transformation can be computed by setting

$$D = \begin{bmatrix} D_1 & 0 & 0 \\ 0 & D_2 & 0 \\ 0 & 0 & D_3 \end{bmatrix}.$$

General expressions for Eq. (8) with respect to l become

$$l_{jk} = \frac{1}{a_{kk}} \left(a_{jk} - \sum_{s=1}^{k-1} l_{js} \cdot l_{ks} \cdot d_k \right) \quad \text{for } j > k \tag{9}$$

$$d_k = a_{kk} - \sum_{s=1}^{k-1} l_{ks}^2 \cdot d_k \tag{10}$$

where j, k, s are indices of the matrices.

A naive implementation of Eq. (9) is straightforward for relatively small matrices. However, when the size of the matrix grows, logic consumption and complexity of the module rise significantly. Thus an alternative approach has been adopted based on the concept introduced in [21]. Given a matrix

$$A_{n-1} = L_{n-1} D_{n-1} A_{n-1}^T, \tag{11}$$

a new matrix can be expressed as

$$A_n = \begin{bmatrix} A_{n-1} & \mathbf{x} \\ \mathbf{x}^T & p \end{bmatrix} = \begin{bmatrix} L_{n-1} & 0 \\ \mathbf{z}^T & 1 \end{bmatrix} \begin{bmatrix} D_{n-1} & 0 \\ 0 & d_n \end{bmatrix} \begin{bmatrix} L_{n-1}^T & \mathbf{z} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} L_{n-1}^T L_{n-1} D_{n-1} & L_{n-1} D_{n-1} \mathbf{z} \\ L_{n-1}^T \mathbf{z}_{n-1}^T D_{n-1} & \mathbf{z}\mathbf{z}_T D_{n-1} + d_n \end{bmatrix}.$$

Equaling the relevant elements of the matrices yields the following set of equations

$$\mathbf{x} = L_{n-1} D_{n-1} \mathbf{z} \tag{12}$$

$$p = \mathbf{z}\mathbf{z}^T D_{n-1} + d_n \tag{13}$$

$$d_n = p - \sum_{k=1}^{n-1} d_k \mathbf{z}_k^2 \tag{14}$$

where Eq. (12) can be presented as

$$L_{n-1} D_{n-1} \mathbf{z} = \mathbf{x} \tag{15}$$

$$L_{n-1} \mathbf{y} = \mathbf{x} \tag{16}$$

$$D_{n-1} \mathbf{z} = \mathbf{y} \tag{17}$$

The solution of Eq. (15) is derived from Eqs. (16) and (17) via substitution.

3) Architecture of the Cholesky decomposition module: The architecture of the proposed Cholesky factorization module is explained by using the 4×4 matrix example used to generate the 5×5 matrix. The structure of the module shown in Figs. 9 and 10 is based on Eqs. (12-17) which are exemplified as follows.

$$\begin{bmatrix} 1 & & & \\ l_{21} & 1 & & \\ l_{31} & l_{32} & 1 & \\ l_{41} & l_{42} & l_{43} & 1 \end{bmatrix} \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & d_3 & 0 \\ 0 & 0 & 0 & d_4 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

$$\begin{bmatrix} 1 & & & \\ l_{21} & 1 & & \\ l_{31} & l_{32} & 1 & \\ l_{41} & l_{42} & l_{43} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

$$\begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & d_3 & 0 \\ 0 & 0 & 0 & d_4 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

We conceptualize two architectures of the hardware module. The first is a strongly pipeline architecture depicted in Fig. 9. It can be utilized for small matrices, but the implementation for larger matrices can be quite problematic. It may serve as an alternative for the second architecture which is more scalable, cf. Fig. 10. The first architecture delivers results every clock cycle but it works properly only for a dedicated and limited matrix size. Multiple uses of the structure would not be straightforward if there is a need to compute 6×6 matrices in contrast to the second architecture which can process matrices of arbitrary sizes. Nevertheless the implementation of the scalable architecture imposes a challenge. The control unit must be designed such that the idle states are minimized. The

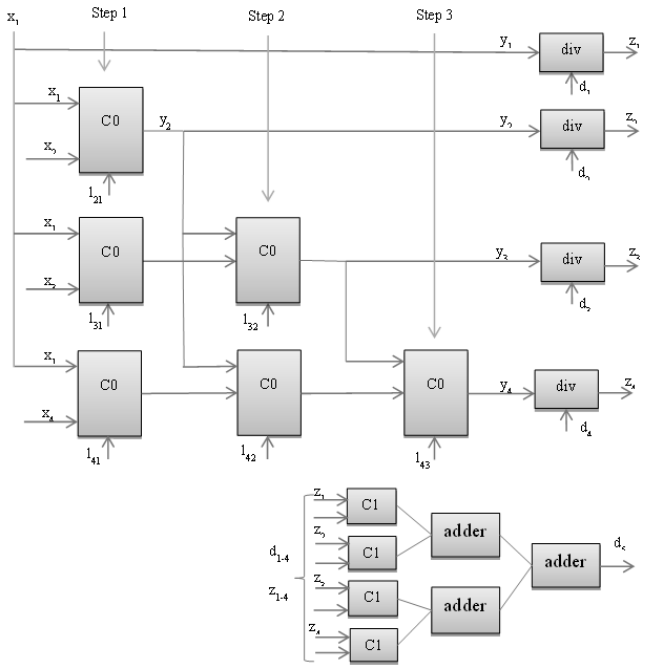


Fig. 9. A pipeline architecture for Cholesky decomposition.

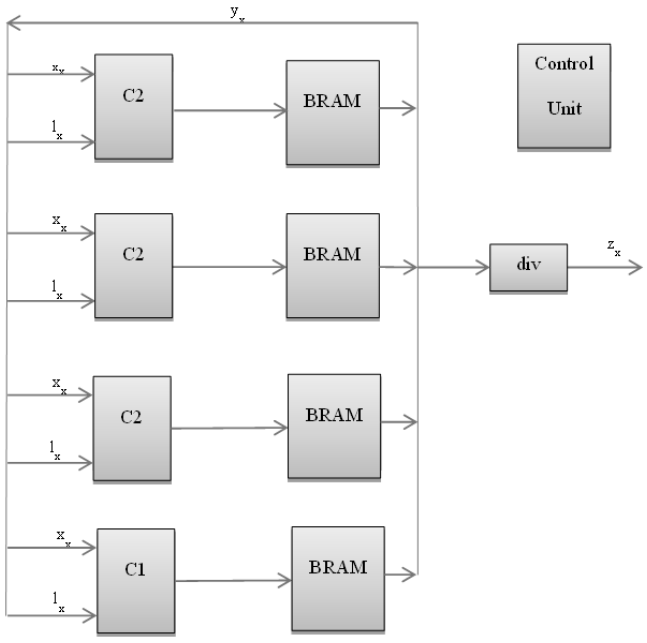


Fig. 10. A scalable architecture for Cholesky decomposition.

TABLE I. THE EXECUTION STEPS OF 5 × 5 MATRIX CALCULATION MODULE.

Step 1	Step 2	Step 3
$y_1 = x_1$	$y_3 = x_{3_{\text{step1}}} - l_{32} y_2$	$y_4 = x_{4_{\text{step2}}} - l_{43} y_3$
$y_2 = x_2 - l_{21} y_1$	$y_4 = y_{4_{\text{step1}}} - l_{42} y_2$	
$y_3 = x_3 - l_{31} y_1$		
$y_4 = x_4 - l_{41} y_1$		

processing phases should overlap according to Table I and Eqs. (12-17). Fig. 11 shows the structure of the linear solvers

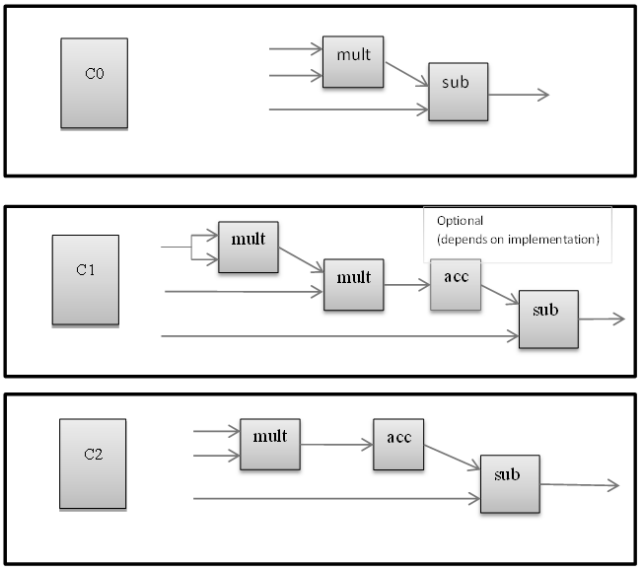


Fig. 11. The internal structure of the C0, C1, C2 building blocks.

in the two architectures.

The algorithm defined by Eqs. (12-14) is based on a gradual decomposition. It is conducted in a row-wise fashion by iteratively solving triangular linear equations starting with the top-left element of the matrix A. As shown in Fig. 10, one triangular linear solver can be used for the whole procedure and subsequently to compute the λ coefficients via Eq. (4). Unlike the calculation of the missing points which is performed multiple times, a variogram matrix is computed just once.

The Cholesky decomposition module is one of those to be implemented on FPGA such as the network units, e.g. router, framer composers and decomposers, quality shaping blocks, and the PCIe components [15]. Due to the scalability of the architecture, it is possible to trade the size of the linear solver for accommodating the other modules in FPGA.

III. Computational Complexity and FPGA Resource Consumption

A. Computational Complexity

There is a tight relationship between the number of the known pixels and the size of the matrix A in Eq. (11). The less pixels are known, i.e. the more pixels to be interpolated, the smaller A becomes, which means less computations involved in its decomposition. On the other hand, a low number of original pixels present in the final picture means a large number of them are to be interpolated. This results in multiple instances of Eqs. (3) and (4) to be solved as the following arguments.

The number of points D_p to be used for calculating the matrix A in Eq. (10) is given by

$$D_p = \sqrt{\frac{D^2}{f_d}} = \frac{D}{\sqrt{f_d}} \tag{18}$$

where D is the size of the original matrix (the length of the original image edge in pixels) and f_d denotes the decimation factor (the number of the dropped points).

The number of CLK cycles required for a single iteration of Eq. (10) as denoted by N_{CLK} is expressed as

$$N_{\text{CLK}} = \frac{1 + (D_p - 1)}{2N_p} = \frac{D_p}{2N_p}$$

where N_p is the number of parallel processing units. The worst case or the longest computation time occurs when only one processing unit in Fig. 10 is implemented.

The number of all the CLK cycles required to calculate a complete matrix A, denoted as $N_{\text{CLK\_complete}}$, is proportional to the size of the variogram matrix formed for a given image D_p. It is given by

$$N_{\text{CLK\_complete}} = D_p N_{\text{CLK}} = \frac{D^2}{2N_p f_p}.$$

Approximately 33,000 clock cycles are needed in the worst case when computing a 256×256 LDL^T matrix as shown in Fig. 12. It is when both the decimation factor f_d and the number of parallel units equal 1. Assuming that the FPGA is clocked at 100 MHz, it takes roughly 330μs to perform the computation.

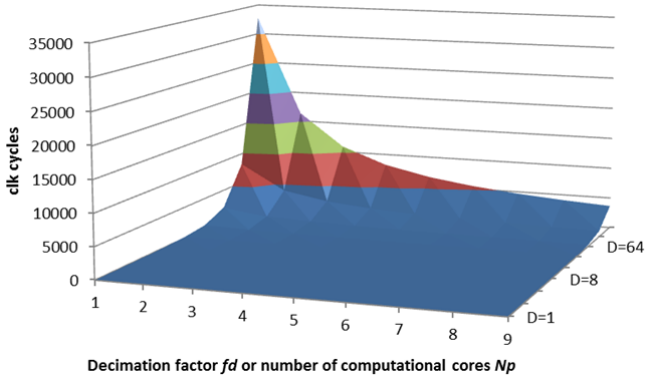


Fig. 12. The number of CLK cycles essential to compute a complete an LDL^T matrix as a function of the decimation factor or the number of computational cores as given by Eq. (18).

The number of the missing points D_m is given by

$$D_m = D^2 - D_p^2 = D^2 \left(1 - \frac{1}{f_d}\right).$$

Interpolating each of the missing points requires performing the following series of matrix operations to generate the λ coefficients:

$$A_n \lambda = \gamma_p \tag{19}$$

$$L_n D_n L_n^T \lambda = \gamma_p \tag{20}$$

$$L_n \mathbf{y} = \gamma_p \tag{21}$$

$$D_n L_n^T \gamma = \mathbf{y} \tag{22}$$

$$D_n \mathbf{z} = \mathbf{y} \tag{23}$$

$$L_n^T \gamma = \mathbf{z}. \tag{24}$$

The computation of Eqs. (22) and (23) can be performed in one step due to the properties of the linear solver. Twice more operations are needed for calculating a missing point than for generating an $L\,D\,L^T$ matrix.

The number of CLK cycles required for a single iteration of λ vector generation, denoted by $N_{\text{CLK\_}\lambda}$, is given by

$$N_{\text{CLK\_}\lambda} = 2\,N_{\text{CLK}} = \frac{D_p}{N_p}.$$

Computing $N_{\text{CLK\_complete\_point}}$, which denotes the number of CLK cycles required to compute all the missing points, follows the equation below:

$$N_{\text{CLK\_complete\_point}} = D_m\,N_{\text{CLK\_}\lambda} = D^3\left(\frac{f_d - 1}{N_p f_d^{3/2}}\right).$$

The computational complexity of the routine for calculating the missing points is $O(D^3)$.

Vector p in Eq. (3) can be calculated in parallel to solve the linear equations for the missing points. The λ coefficients being generated can be used right away in Eq. (4). It means that, as the λ coefficients are generated for a given point, its value is also computed. This operation can be implemented as a multiplier accumulator block or a single DSP48E. Therefore it is not accounted for in computing $N_{\text{CLK\_complete\_point}}$.

For $D = 256$ pixels and a single computational core, the number of clock cycles needed to compute all the missing points reaches 5×106 CLK cycles (see Fig. 13). The computational time in this case is roughly 50ms, assuming that the FPGA is clocked at 100 MHz. It is far beyond the system latency limit. This is the worst-case assumption, which means that all the 256×256 points are to be interpolated. In practice the case when the smallest f_d is N^2 as in Fig. 1 never occurs.

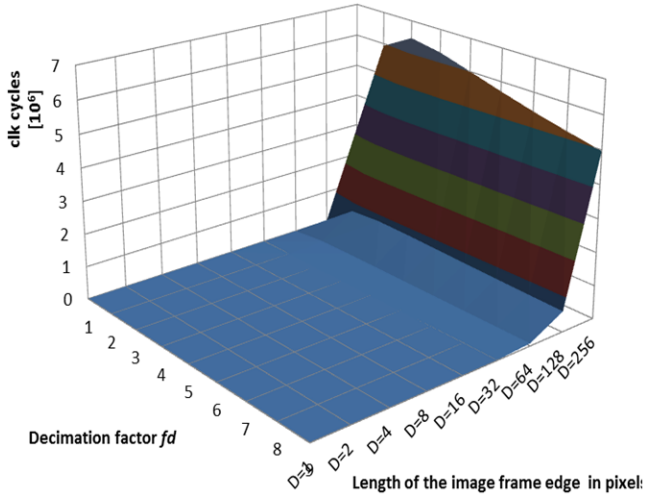


Fig. 13. The number of CLK cycles essential for interpolating the missing points in an input image.

B. Resource Consumption

The estimated FPGA resource consumption in a Xilinx Virtex-6 XC6VLX240T for the building blocks in Figs. 9 and 10 is presented in Table II. A Xilinx ML-605 board equipped with a Virtex-6 XC6VLX240T is the chosen FPGA platform for the Kriging module implementation. The FPGA contains 241,152 logic cells, 150,720 LUTs, 301,440 FFs, 768 DSP48E1 and 832 18Kb BRAM memories [19]. Thus a single Virtex-6 XC6VLX240T can accommodate $150720/217 = 694$ C2 blocks, the basic building cores of the linear solver. The limitation here is the number of LUT memories. The estimate does not account for available DSP48 blocks.

TABLE II. RESOURCE CONSUMPTION OF THE LINEAR SOLVER BUILDING BLOCKS IN FIGS. 9 AND 10.

Module	#LUT	#FF	#DSP slices	#BRAM
C0	123	110	0	0
C0 (DSP)	0	0	1	0
C1	332	343	0	0
C2	217	233	0	0
C2 (DSP)	70	91	1	0

If all 694 C2 modules are exhausted to compute the 256×256 missing points, the total processing time would drop to $50\text{ms}/694 \approx 0,072\text{ms}$. Unfortunately only a part of all the FPGA resources can be devoted to KIM unless there is a separate FPGA dedicated only for its implementation. Fig. 14 presents a block diagram of an exemplary KIM which is capable of processing three windows in parallel [11].

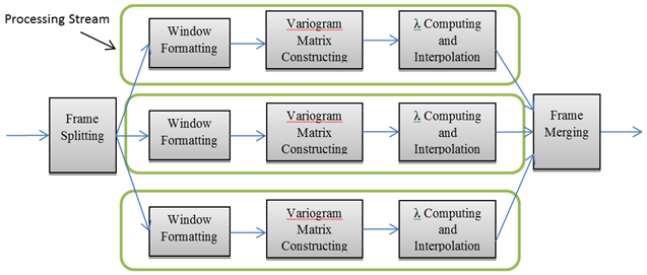


Fig. 14. A block diagram of a Kriging interpolation module.

The module works as follows. The Frame Splitting unit divides the image in window size chunks and sends them to the Window Formatting blocks. The Window Formatting unit builds a matrix of the known pixels locations and the set of vectors of unknown points. These will subsequently be fed into the Variogram Matrix Constructing module. In most cases it is just a LUT memory. The module generates a variogram matrix and passes the results to the λ Computing and Interpolation Module. As the architecture depicted in Fig. 14 is scalable, an arbitrary number of processing streams may be chosen. The optimal performance is achieved when the number of the processing streams is equal to the number of windows [11].

Let us assume that a single processing stream is used to process 256×256 images. It absorbs 25% of all the resources of Virtex-6 XC6VLX240T. Note here that 173 units of C2 are used, cf. Fig. 11. In that case the generation of the $A = L\,D\,L^T$ matrix takes $330\mu\text{s}/173 \approx 1.9\mu\text{s}$ and computing all the missing points lasts $4 \times 0,072\text{ms} = 0,288\text{ms}$. The overall execution time is expected to be roughly 0,3ms in such a case. The assumption of 25% logic utilization is reasonable because it leaves space for implementing the remaining modules, e.g. PCIe, DDR3 memory controller, DMA and Microblaze.

IV. Conclusion and future direction

The Kriging interpolation module (KIM) is a part of the quality shaping scheme which is considered as one of key components of the future networked collaboration system. Controlled dropping of packets leads to graceful quality degradation. This is achieved provided an effective interpolation mechanism is implemented for very demanding situations. Therefore FPGA realization of kriging is critical and requires special effort to meet a very strict EED of 11ms. A scalable KIM architecture is proposed and its implementation feasibility is analyzed with respect to the DMP system. It poses several challenges to be addressed as future work, such as the implementation of the variogram fitting routine and the impact of using fixed variogram matrix on the quality of the interpolation. The modularity of the proposed architecture makes its future modification or extension straightforward. The latency introduced by the module is less than 1ms for a 256×256 image.

References

[1] H. Berge, M. Panggabean, and L.A. Rønningen, "Modelling video-quality shaping with interpolation and frame-drop patterns," in *Proc. 23rd Norsk informatikkonferanse (NIK)*, 2010, pp. 132–143.

[2] D. Besiris, V. Tsagaris, N. Fragoulis, and C. Theoharatos, "An FPGA-based hardware implementation of configurable pixel-level color image fusion," *IEEE Trans. Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 362–373, 2005.

[3] C. Chafe, M. Gurevich, G. Leslie, and S. Tyan, "Effect of time delay on ensemble accuracy," in *Proc. Int'l Symp. Musical Acoustics*, 2004.

[4] T. Cheng, D. Li, and Q. Wang, "On parallelizing universal kriging interpolation based on OpenMP," in *Proc. 9th Int'l Symp. Distributed Computing and Applications to Business Engineering and Science (DCABES)*, 2010, pp. 36–39.

[5] E. Decenciere, C. de Fouquet, and F. Meyer, "Applications of kriging to image sequence coding," *Signal Processing: Image Communications*, vol. 13, no. 3, pp. 227–249, 1998.

[6] F. He, J. Fang, and W. Zou, "An effective method for interpolation," in *Proc. 19th Int'l Conf. Geoinformatics*, 2011, pp.1–6.

[7] T. Hengl, *A Practical Guide to Geostatistical Mapping*, Joint Research Centre Institute for Environment and Sustainability, European Commission, 2009.

[8] M. Li and L. Dong, "Visualization three-dimensional geological modeling using CUDA," in *Proc. 6th Int'l Conf. Image and Graphics (ICIG)*, 2011, pp. 852–857.

[9] O. Maslennikow, P. Ratuszniak, and A. Sergyienko, "Implementation of Cholesky LLT-decomposition algorithm in FPGA-based rational fraction parallel processor," in *Proc. 14th Int'l Conf. Mixed Design of Integrated Circuits and Systems*, 2007, pp. 287–292.

[10] A. Panagiotopoulou and V. Anastassopoulos, "Super-resolution image reconstruction employing Kriging interpolation technique," in *Proc. 14th Int'l Workshop on Systems, Signals and Image Processing (IWSSIP)*, 2008, pp.114–147.

[11] M. Panggabean, Ö. Tamer, and L.A. Rønningen, "Parallel image transmission and compression using windowed kriging interpolation," in *Proc. 10th IEEE Symp. Signal Processing and Information Technology (ISSPIT)*, 2010, pp. 315–320.

[12] M. Panggabean and L.A. Rønningen, "Chroma interpolation using windowed kriging for color-image compression-by-network with guaranteed delay," in *Proc. 17th Int'l Conf. Digital Signal Processing (DSP)*, 2011, pp.1–6.

[13] M. Panggabean and L.A. Rønningen, "Parameterization of windowed kriging for compression-by-network of natural images," in *Proc. 7th Int'l Symp. Image and Signal Processing and Analysis (ISPA)*, 2011, pp. 373–378.

[14] J. Ruiz-Alzola, C. Alberola-Lopez, and C.F. Westin, "Kriging Filters for Multidimensional Signal Processing," *Signal Processing*, vol. 85, no. 2, pp. 413–439, 2005.

[15] L.A. Rønningen, *The DMP System and Physical Architecture*, Technical Report, Department of Telematics, Norwegian University of Science and Technology, 2007.

[16] L.A. Rønningen, M. Panggabean, and Ö. Tamer, "Toward futuristic near-natural collaborations on Distributed Multimedia Plays architecture," in *Proc. 10th IEEE Symp. Signal Processing and Information Technology (ISSPIT)*, 2010, pp.102–107.

[17] J. Strzelczyk, S. Porzycka, and A. Lesniak, "Analysis of ground deformations based on parallel geostatistical computations of PSInSAR data," in *Proc. 17th Int'l Conf. Geoinformatics*, 2009, pp.1–6.

[18] Xilinx, http://www.xilinx.com/products/boards-and-kits/EK-V6-ML605-G.htm, (2012).

[19] Xilinx, http://www.xilinx.com/support/documentation/data_sheets/ds150.pdf, (2012).

[20] Ch. Xuezheng, K. Benkrid, and J. Thompson, "Rapid prototyping of an improved Cholesky decomposition based MIMO detector on FPGAs," in *Proc. NASA/ESA Conf. Adaptive Hardware and Systems*, 2009, pp.369–375.

[21] D. Yang, G. Peterson, and H. Li, "High performance reconfigurable computing for Cholesky decomposition," in *Proc. Symp. Application Accelerators in High Performance Computing (SAAHPC)*, 2009.

[22] Z. Yang, B. Yu, W. Wu, K. Nahrstedt, R. Diankov, and R. Bajscy, "A study of collaborative dancing in tele-immersive environments," in *Proc. 8th Int'l Symp. Multimedia*, 2006, pp. 177–184.

Ultrafast Scalable Embedded DCT Image Coding for Tele-immersive Delay-Sensitive Collaboration

Mauritz Panggabean, Maciej Wielgosz, Harald Øverby, and Leif Arne Rønningen
Department of Telematics (ITEM)
Norwegian University of Science and Technology (NTNU)
N-7491, Trondheim, Norway

***Abstract*—A delay-sensitive, real-time, tele-immersive collaboration for the future requires much lower end-to-end delay (EED) for good synchronization than that for existing teleconference systems. Hence, the maximum EED must be guaranteed, and the visual-quality degradation must be graceful. Distributed Multimedia Plays (DMP) architecture addresses the envisioned collaboration and the challenges. We propose a DCT-based, embedded, ultrafast, quality scalable image-compression scheme for the collaboration on the DMP architecture. A parallel FPGA implementation is also designed to show the technical feasibility.**

I. Introduction

Figure 1 shows a simple example of the envisioned collaboration. A and B engage each other in a real-time delay-sensitive communication. They are both a source and a receiver, whereas C only receives data from them. As EED is not critical for C, C can use video-streaming technologies over the Internet. The capacity in the multihop links between A, B and C varies because other users outside the collaboration also use them. Moreover, the target quality of experience (QoE) is so high that it closely approximates reality, i.e. near-natural.

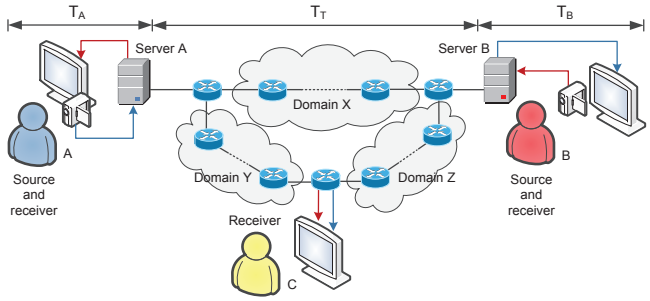


Fig. 1: A simple example of the futuristic collaboration.

The collaboration environment is an immersive collaboration space (CS) more advanced than the CAVE [5]. To achieve the near-natural QoE, each surface in the CS is tiled with autostereoscopic multiview 3D displays with arrays of high-end cameras, microphones, and speakers. At high frame rate, the video traffic from a CS is several orders of magnitude higher than that in typical videoconferencing. It is reduced by *segmenting* only the important objects in the video, such as the faces and bodies of the performers.

A maximum EED for good synchronization between A and B must be guaranteed. Some studies show that the optimal EED for synchronizing rhythmic clapping hands from different places is 11.5ms [4]. Longer delays will produce increasingly severe tempo deceleration while shorter ones yield a modest yet surprising acceleration. Since musical instruments such as percussion are rhythmically very similar to clapping hands, percussion musicians who collaborate from remote places require the same EED for synchronization. It also applies to collaborative dancing because dancers perform based on visual cues from each other [40]. Other cases include collaborative singing and remote conducting [18].

An EED consists of delays due to propagation, transmission, and signal processing. Propagation delay is caused by physical distances, and transmission delay depends on link capacity, queueing delay, and computations at the network nodes. The latter is the electronic bottleneck that limits the achievable capacity of a network [27]. Less capacity in the network causes congestion and increases queuing delay. Instead of multipath transmission, single path is assumed to simplify routing delay. Exploiting temporal redundancy when encoding video data gives better quality but increases encoding delay. *Intraframe* video encoding is, therefore, preferred as shown by a recent experiment that uses JPEG [10]. Since we pursue *very low latency for encoding and decoding* in the order of μs per frame, *parallel* computation must be used as much as possible.

The Distributed Multimedia Plays (DMP) architecture has been proposed to facilitate the envisioned collaboration [17] with the idea that maximum EED is guaranteed if each network node guarantees that the local delay never exceeds its maximum value. This value and the propagation delay can be estimated prior to packet transmission. Because the routers and switches in DMP have advanced functionalities to guarantee Quality of Service (QoS), DMP belongs to the network-centric approach rather than the end-system-based approach [37].

The idea has three important implications. First, a DMP network node must be able to drop parts of the video packets deliberately whenever necessary to guarantee its local delay. The dropping must be fast, and the buffer size must be optimal. Determining the latter is not the goal of this work.

Second, the packet dropping to guarantee graceful video-quality (VQ) degradation must be conducted intelligently. The video contents in the packets must be arranged and transmitted in decreasing order of the importance to VQ. The less important the contents in a packet, the higher the dropping priority. Packets that contain very essential contents, however, must never be dropped. This leads to the property of *quality*

scalability in the wanted image-compression scheme.

Third, fast packet dropping means that it occurs in compressed domain. By providing information necessary for this in the bitstream, the cycle of decoding, dropping, and re-encoding at a node is avoided. This and the second implication mean that the bitstream can be truncated at any point to yield the reconstructed image at a lower bitrate. The quality at the final rate after dropping should be the same with that if it is encoded directly at that rate, i.e. *embedded coding* [22].

The objective of this work is to design an image-compression scheme that has all the properties aforementioned: ultrafast, embedded, quality scalable, fully parallelized, and supporting the processing of segmented objects with arbitrary shapes. Note that we do not pursue better coding performance than that of non-scalable image coding standards because it is unfair and irrelevant. The envisioned collaboration allows for sub-optimal VQ as the price for guaranteeing maximum EED as long as the VQ is gracefully degraded. Tradeoff is normal in image/video coding. For instance, H.264/MPEG-4 AVC [11] and x264 [39] provide profiles and presets to meet various priorities such as low complexity or high performance.

This paper is structured as follows. Section II details the proposed image-compression technique. Experimental results follow in Section III with discussion and analysis. Section IV discusses the complexity of the algorithms for implementation on field-programmable gate array (FPGA). Section V concludes the paper with summary and further ideas.

II. The Proposed Image-Compression Scheme

The DMP approach resembles the concept of layered coding such as in scalable video coding (SVC) [15], [19], [25] and JPEG 2000 [32]–[34]. SVC achieves temporal, spatial, and quality scalability by removing parts of the video bitstream to adapt it to different end-users' preferences and varying terminal capabilities or network conditions. Proposed to supersede JPEG, JPEG 2000 is an image compression standard and coding system based on wavelet transform. Some of the improvements over JPEG are as follows: superior compression performance, multiple resolution representation; progressive transmission by pixel and resolution accuracy; spatial, quality and channel scalability; support of lossless and lossy compression; embedded coding; facilitated processing of regions of interest; error resilience. Consequently, they make JPEG 2000 more complex and computationally demanding.

The properties aforementioned make the proposed image-compression technique (Fig. 2) somewhat different from the existing ones. For example, the quantization, a key step such as in JPEG image compression (Fig. 3), is the principal cause for the loss of information, while the loss in DMP is due to the deliberate packet dropping at network nodes, i.e. the proposed scheme has no such quantization. Moreover, the techniques optimize bandwidth utilization by aiming for the best VQ at a given bitrate with no guarantee over maximum EED.

A. Block Ranking and Transform

The encoder of the proposed scheme consists of three major steps: block ranking, transform, and entropy coding (variable length encoding, VLE, and run-length encoding, RLE). After

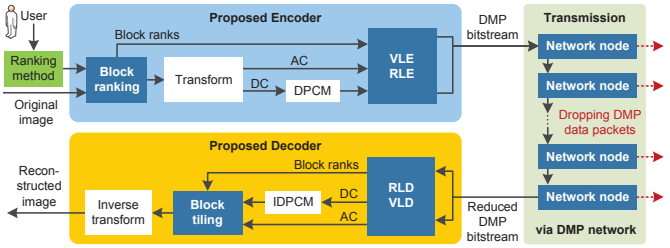


Fig. 2: The proposed image-compression technique.

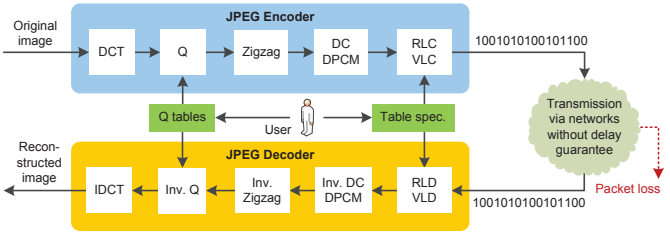


Fig. 3: Block diagram of JPEG image compression.

an input picture is divided into $N \times N$ blocks and the color space is converted to YCbCr, the block ranking automatically classifies each block into one of several ranks according to how important the block contents are to VQ. The importance of a block is indicated by the level of distortion to human perception when the binary representation of the block content becomes less precise, e.g. via quantization or packet dropping. The blocks are independent from each other, and thus can be processed concurrently. In this work $N = 8$ pixels, and users can define their own ranking method.

For the transform, two dimensional DCT (2D-DCT) [1] is selected for two main reasons. First, it is widely used because of the excellent energy compaction. Second, many fast hardware (HW) implementations of 2D-DCT have been reported, e.g. in [28]. The most recent work closest to ours is that by van der Vleuten et al. [35], which incorporates quality scalability to JPEG by encoding the DCT coefficients bit-plane by bit-plane, starting at the most significant one. Although the performance is similar to that of JPEG without quantization or entropy coding, the algorithm, particularly the scan order, must be adapted to each image. Our scheme is agnostic to the input image.

The block ranking can be applied before or after the block transform. The first only has 64 pixel values of the block luma available for analysis and ranking, whereas 64 DCT coefficients are additionally present in the latter. We choose the first option because various statistical properties of pixel values have been used for content classifications in images [6], [41]. Furthermore, luma values are integers, but DCT coefficients use floating points. Therefore, computing pixel values requires less resources and time than if DCT coefficients are added to the computation. Moreover, DCT coefficients in natural images are more complex to use for classification purposes [24].

The statistical measures for ranking the blocks must be highly accurate and fast to compute. We use the entropy E, which measures the amount of information and uncertainty

contained in data [21]. For a grayscale image with N unique pixel values, it characterizes the texture therein as

$$E = -\sum_{i=1}^{N} p_i \log_2 p_i$$

where p_i is the probability of the ith pixel value from the histogram counts. The block entropy BE rises when the frequency content of the block increases.

Shown in Fig. 4 for LENA image using 8×8 blocks, the BE values are between 2 and 6 in all images tested (Fig. 5). Since the colors correspond well with human perception, BE is a good indicator of the frequency content in a block. The constant range of BE can be used to define the thresholds for arbitrary number of block ranks for dropping. We use the following four block ranks with the ranges: low ($\lceil BE \rceil \leq T_L$), low-medium ($T_L < \lceil BE \rceil \leq T_{LM}$), medium-high ($T_{LM} < \lceil BE \rceil \leq T_{MH}$), and high ($\lceil BE \rceil > T_{MH}$), where T_L, T_{LM} and T_{MH} are thresholds in positive integers.

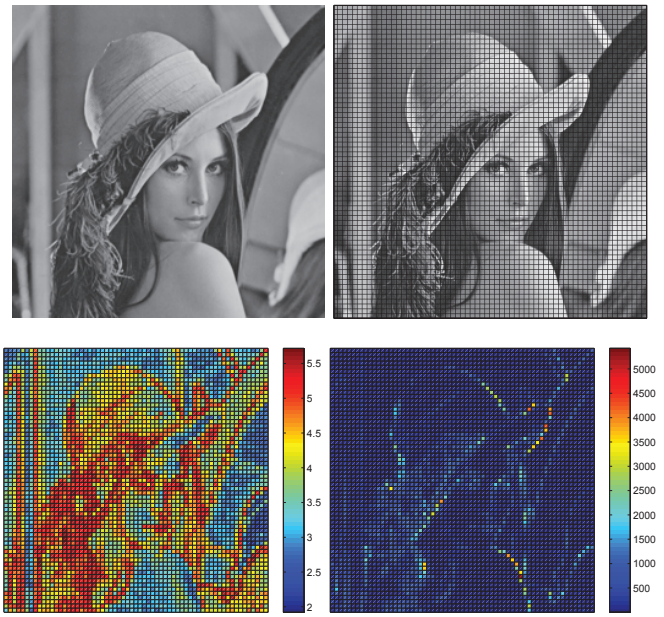


Fig. 4: Clockwise from top left: the original LENA image, the tiled 8×8 blocks, the BV, and the BE.

The criterion for the high-frequency rank is not very accurate because some of the blocks are grouped into the medium-high rank. It leads to the use of block variance BV to improve the block-ranking accuracy. The variance of a grayscale image with M pixel values is given by

$$V = \frac{1}{M-1} \sum_{i=1}^{M} (x_i - \hat{x})^2$$

where x_i denotes the intensity value of the ith pixel, and $\hat{x}$ is the average of all the pixel values. In the proposed block-ranking algorithm (Algorithm 1), $\lceil x \rceil$ rounds the scalar x to the nearest integer towards plus infinity, $T_V = 1, T_L = 3$, and $T_{LM} = 4$. The resulting BV values for LENA image are shown in Fig. 4. The BE and BV are not only fast to compute in HW (section IV), but also correspond well with human perception (section III).

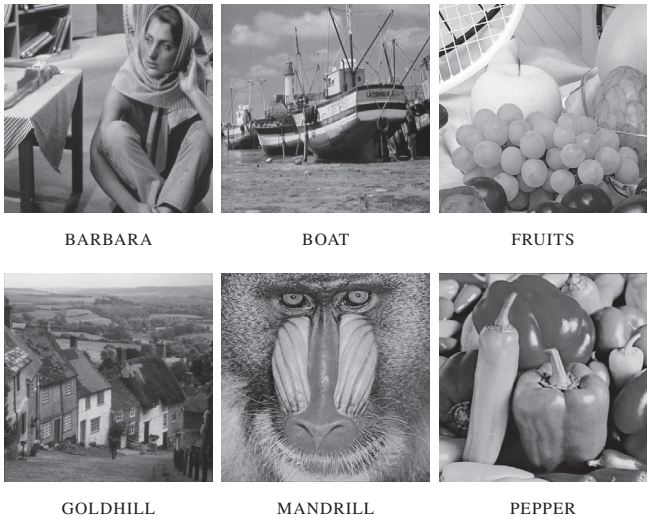


Fig. 5: The test images besides LENA.

Algorithm 1 Proposed algorithm for block ranking

```
if ⌈BV/100⌉ ≤ T_V then
    if ⌈BE⌉ ≤ T_L then
        Rank 4: low frequency (blue)
    else if T_L < ⌈BE⌉ ≤ T_LM then
        Rank 3: low-medium frequency (green)
    else if ⌈BE⌉ > T_LM then
        Rank 2: medium-high frequency (yellow)
    end if
else
    Rank 1: high frequency (red)
end if
```

Encoded and included in the bitstream as side information, the produced block ranks must never be lost because it will jeopardize the image reconstruction from the transmitted packets at the receiver. They are also used in structuring the encoded DCT coefficients into the data packets to enable packet dropping in the compressed domain.

After the block ranking, 2D-DCT is applied to each of the luma block independently, and it produces 64 DCT coefficients per block, which comprise one DC coefficient and 63 AC coefficients. As the DC coefficient contains the average value of the block, it is essential for reconstruction and must not be dropped. The location of an AC coefficient in a block indicates the importance. The only information loss in the proposed encoder, rounding the values to the nearest integers reduces the precision for faster computation with less memory use.

B. Universal Codes for Entropy Coding

The distribution of the rounded DCT coefficients is key in encoding them losslessy and efficiently. Fig. 6 depicts the empirical probability density functions (PDFs) of the rounded AC coefficients between -10 and 10 from DCT and Walsh-Hadamard transform (WHT) for comparison. They include more than 99% of all the coefficients for WHT in all the test images, but it is between 80% and 99% for the DCT. Note the symmetry around zeros in the PDFs. The quality of the

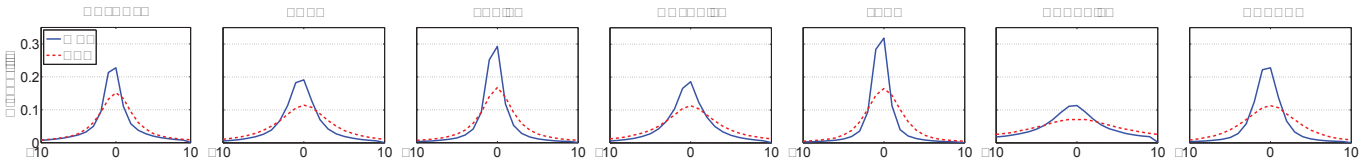


Fig. 6: The empirical PDFs of the AC coefficients from DCT and WHT for the test images.

reconstructed images using DCT in peak-signal-to-noise ratio (PSNR) is always a few dB higher than that for WHT due to the rounding of the DCT coefficients.

The probability of zeros is always less than 0.20 for DCT and slightly higher for WHT. Producing shorter average length of codewords with higher probability of zeros [23], the Huffman code is not efficient for this case. Moreover, the distribution of the DCT coefficients could not be known beforehand. The Huffman code is favored when more zeros are present in the high-frequency coefficients caused by stronger quantization.

The dropping of the DC coefficients starting from the least important implies that RLE is not suitable for this work because it introduces more dependencies in the resulting bitsream. The run lengths of the coefficients can also be very short because of no quantization. This is a challenge because RLE can increase performance in lossless coding.

Furthermore, coding techniques that can only be decoded when the bitstream is complete, such as the Burrows-Wheeler transform (BWT) [3], are also not suitable. It is, in fact, impossible because the received stream at the end are truncated due to dropping. Nevertheless, they are useful for encoding the block ranks and the rounded differences after applying differential pulse-code modulation (DPCM) to the DC coefficients.

Excellent texts such as [23] comprehensively discuss and compare various coding techniques available. They lead us to the use of universal codes (UCs) for entropy coding for the applicability regardless of the data distribution. We propose using the Fraenkel and Klein C^1 Fibonacci code [8] (FK_1) based on the comparison of well-known UCs in [7]. The recurrence relation $F(i) = F(i-1)+F(i-2)$ with seed values $F(0) = 0$ and $F(1) = 1$ defines the sequence $F(i)$ of the famous Fibonacci numbers. The Zeckendorf's theorem states that any integer can be formed as the sum of Fibonacci numbers [29]. Thus, for a positive integer number n, if $d_0, d_1, ..., d_k$ represent n, then we have $n = \sum_{i=0}^{k-1} d_i F_{i+2}$ and $d_k = d_{k-1} = 1$ where F_i is the ith Fibonacci number. A Zeckendorf representation $Z(n)$ is coded by writing a binary vector with a 1 wherever that Fibonacci number is included, but F_1 is omitted due to redundancy. For example, since $19 = 13(F_7) + 5(F_5) + 1(F_2)$, it means $19 = (1\times13)+(0\times8)+(1\times5)+(0\times3)+(0\times2)+(1\times1)$ which gives $Z(19) = 101001$.

A very important property of $Z(n)$ is that two adjacent 1's never occur. Therefore, the FK_1 code produces $FK(n)$ by writing $Z(n)$ in the reverse order and appending another 1 as a terminating comma; hence, $FK(19) = 1001011$. Decoding n from $FK(n)$ is straightforward and only involves additions, making it fast for HW implementation. Using the code for signed integers is possible after *bijection*, i.e. mapping the real values in signed integers into symbols in positive values. Table I shows the FK_1 codewords of the symbols n from applying bijection to the real values x.

TABLE I: Some examples of $FK_1(n)$ for symbols from real values after bijection

x	n	$FK(n)$	x	n	$FK(n)$
0	1	11	3	6	10011
1	2	011	-3	7	01011
-1	3	0011	4	8	000011
2	4	1011	-4	9	100011
-2	5	00011	...	...	...

The FK_1 code offers several advantages [7]. First, unlike using adaptive parametrized codes, storing tables of codewords in the network nodes for packet dropping is unnecessary; hence, more efficient use of resources. Second, using two 1's as the delimiter between consecutive coded symbols gives more robustness against transmission errors than table-based codes such as the Huffman code. Third, because of the universal codewords, simply reading from lookup tables (LUTs) allows fast encoding and decoding. Fourth, the memory allocated for the LUTs is also very small (section IV). Fifth, no prefix code used also means higher efficiency.

The encoding strategy for the DCT coefficients and the block ranks is proposed as follows. First, the block ranks are encoded in the raster fashion using BWT because only four integer symbols are used, which saves around 18% than using 2-bit binary encoding. Using run lengths gives very little gain, merely around 0.01 kilobytes. The BWT is currently the best lossless compression technique, especially for text, with fast implementations available [23] such as the *bzip2* [20].

Second, the AC coefficients are processed and encoded separately from the DC. Prior to encoding, DPCM and bijection are applied to the signed coefficients. The resulting symbols for the DC coefficients are then encoded into binary string using the FK_1 code. The bitstream from the block ranks and the DC coefficients must not be dropped.

Third, the 63 AC coefficients from each block are grouped into 63 series according to their position, following the zigzag direction as used in JPEG. The symbols after bijection are encoded in parallel using the FK_1 code. The symmetry of their distribution (Fig. 6) motivates the use of bijection. The coefficients starting from the most top-left block are transmitted first in the raster fashion, and the series are sent according to their series number.

If the probability of zeros in the resulting binary string is very high, i.e. higher than 0.9, the run lengths of ones and zeros can be encoded further, for example, using the Golomb codes [9]. In all the test images, however, the probability is

only around 0.7. The bitstream from the Golomb code must be decoded before dropping. Since the reduced bitstream after dropping must be encoded again using the Golomb code, processing time at the network nodes increases.

C. Data Structure and Packet Format

The data structure also affects the coding. The proposed data structure depicted in Fig. 7 can be used to arrange the blocks, ranks and coefficients for encoding. Since all the AC coefficients of the same rank and index are grouped together, dropping them as a group when necessary is straightforward. Packet dropping is fast because checking each block rank for dropping is not needed. Moreover, since the bitstream of the group is long, more compression gain can be achieved using Golomb codes on the run lengths of the zeros. The proposed HW design and implementation of the DMP network node also obtain higher throughput with longer input bitstream. This area opens many interesting questions for future work. For this work we use fixed-length packets.

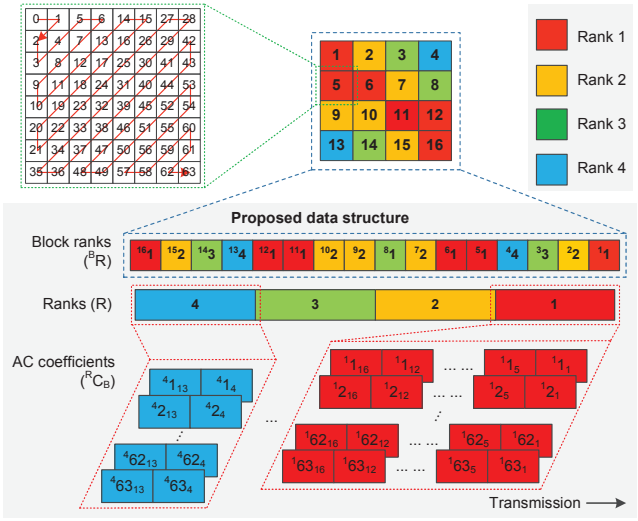


Fig. 7: Proposed data structure for packetization and transmission of the encoded blocks, ranks, and DCT coefficients.

III. Results and Discussion

Seven standard grayscale test images are used in the experiments. The results are produced using MATLAB, and the *bzip2* codec [20] is used for the BWT-based compression. All images exhibited in this section should be seen with magnification on screen for the best perceptual quality.

The first two images in Fig. 8 depict the block maps of LENA image using only entropy and that using the proposed block-ranking method. Both maps have the same blue and green blocks, but not those in the other two colors. There are more red blocks in the second image, for example on the edges and in the area of the fur. This illustrates the importance of the block variance in ranking the blocks. It produces more red blocks because they are more sensitive to visual distortion.

Fig. 8 also displays four sets of areas according to the four ranks produced by the proposed ranking method. They show the high classification accuracy of the proposed ranking

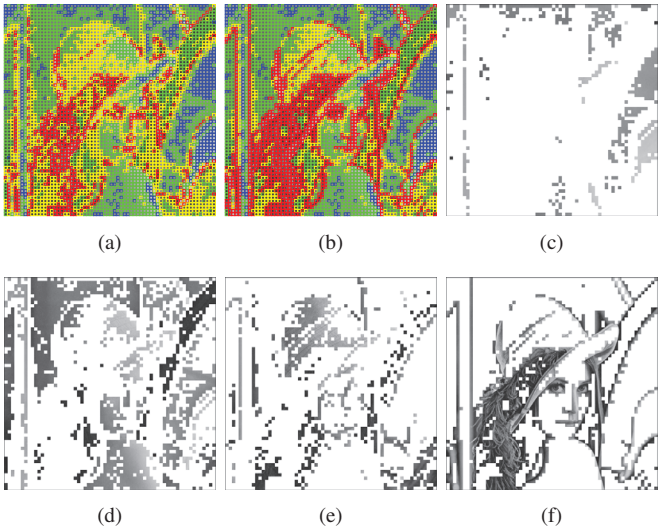


Fig. 8: The four-rank block map of LENA image using only entropy (a) and that using the proposed ranking method (b). The image is decomposed into five ranks as follows (with decreasing dropping priority): low frequency in blue (c), low-medium in green (d), medium-high in yellow (e), and high in red (f). Borders are added for better view.

method. Different techniques to classify the contents can be employed, for example using edge and texture detection to detect the edges and the textured areas. This idea, however, is not necessary because the blocks containing them can be successfully categorized as those in red by the proposed method.

The distribution of the four block ranks in the test images as shown in Fig. 9 indicates that all the rank maps of the images correspond well with human perception. It can be checked by visually comparing the distribution with the original images in Fig. 5. Rank-1 blocks are dominant in BARBARA, BOAT, GOLDHILL, and especially MANDRILL due to many textured areas present therein. In the other images, the portions of the blocks of Rank 1 and Rank 3 are almost the same because of the flat areas (Rank 3) and the edges (Rank 1).

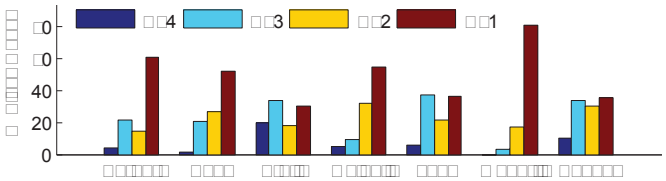
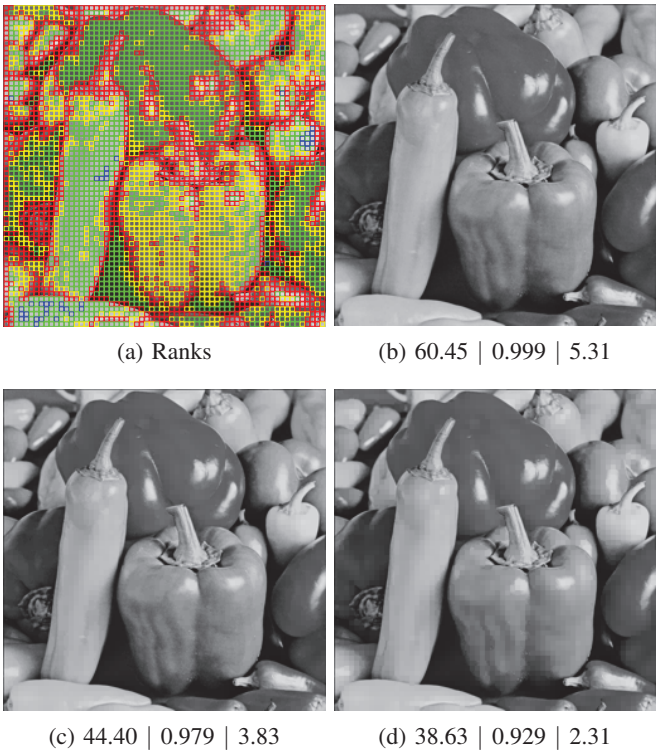


Fig. 9: The distribution of the four ranks in the test images.

Some examples of the rank map and the reconstructed images are shown in Fig. 10 from PEPPER image. Fig. 10 (b) is reconstructed from the received bitstream without the DC coefficients of Rank 4 because they have been dropped completely. When the network capacity is reduced, the nodes start dropping the AC coefficients of Rank 3 beginning from those of the 63th index. When all the AC coefficients of Rank 3 have been dropped, the resulting image quality is shown in



(a) Ranks (b) 60.45 | 0.999 | 5.31

(c) 44.40 | 0.979 | 3.83 (d) 38.63 | 0.929 | 2.31

Fig. 10: The examples from PEPPER image: the rank map with entropy and variance (a); the images reconstructed without the DCT coefficients from Rank 4 (b), from Ranks 4 and 3 (c), and from Ranks 4, 3, and 2 (d). The PSNR (dB), MSSIM and bitrate (bpp) are provided underneath.

Fig. 10 (c). The dropping is continued until the worst quality is achieved by reconstructing only from the DC coefficients of all ranks (Fig. 10 (d)). The results are accompanied with the bitrate in bits per pixel (bpp) as well as the PSNR and the mean structure-similitary index (MSSIM) [30] as the objective VQ metrics. Furthermore, the rate-distortion plots using the two quality measures for the test images are shown in Fig. 11. The plots from JPEG are also provided for comparison.

The performance of the proposed scheme can be improved by using deblocking filter as post-processing step at the receiver. An important role for estimating the bitrate, this step is conducted at the source in many image/video coding techniques. Other possible improvements include intra-prediction, which is not considered here because it creates dependencies between the adjacent blocks and increase the complexity. Note again that we do not aim at better coding performance than that of JPEG or even JPEG 2000.

Fig.10 shows that the reconstructed images suffer from blocking artifacts that occur only at the blocks which AC coefficients are dropped. The artifact, however, is different from the typical blocking artifact in JPEG because the latter occupies much larger areas consisting of many blocks. The distortion is called *pixelation artifact* due to the resemblance to it. The encoded rank maps in the side information of the bitstream plays another important function; they inform the receiver of the exact locations of the pixelated blocks. Thus those blocks can be directly restored without searching their locations as in typical deblocking algorithms.

For the worst distortion because all the AC coefficients are discarded, a fast depixelization algorithm is proposed in Algorithm 2 which refers to Fig. 12. Fig. 13 shows some examples of the depixelization for FRUIT and PEPPER images with PSNR and MSSIM values. The algorithm successfully restores the flattened blocks to be more appealing to human perception.

The blocks of Rank 2, 3 and 4 can be reconstructed only from the DC coefficients without pixelated blocks because they can be repaired fast using Algorithm 2 (Fig. 13). In fact, Rank-1 blocks with strong textures and no edges can be made free from pixelation. Nevertheless, the artifacts are still visible in Rank-1 blocks with edges even after depixelization. Repairing the pixelated edges can benefit from more advanced techniques such as in [13].

Fig. 14 (left) shows a collaborating person as a segmented object from a video frame of an HD test video sequence from [26]. It is expected to be produced by the cameras on a surface of a CS. Applying the proposed block ranking algorithm produces the blocks in Fig. 14 (right). The blocks of the background can be assigned an additional rank, for

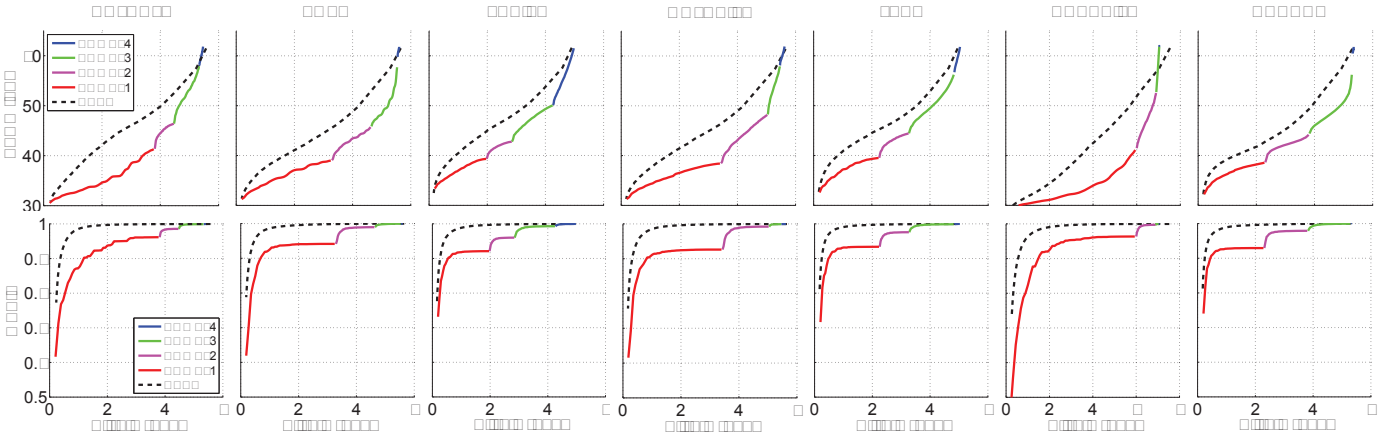


Fig. 11: The rate-distortion plots of PSNR (top) and MSSIM (bottom) against bitrate in bits per pixel (bpp).

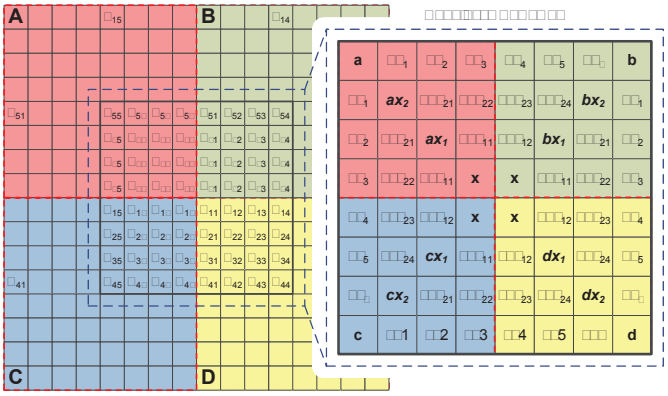


Fig. 12: The proposed depixelization in Algorithm 2.

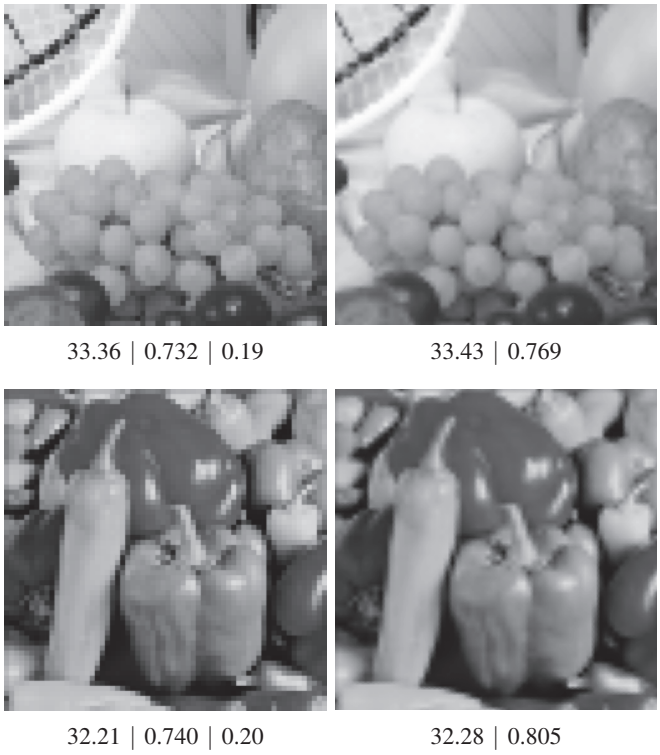
Algorithm 2 A depixelization algorithm for the worst distortion

Reference elements of the subblocks A, B, C and D
$X \leftarrow (A_{88} + B_{81} + C_{18} + D_{11})/4$
$a \leftarrow A_{55}$, $b \leftarrow B_{54}$, $c \leftarrow C_{45}$, $d \leftarrow D_{44}$
for Block $M = \{A, B, C, D\}$ **do**
  $\{m, mx_1, mx_2, X\} \leftarrow \text{LI}(m, X, 2)$
end for

Non-reference elements of the subblocks A, B, C and D
if Rank of block $A \geq T_R$ and $C_A \leq T_C$ **then**
  if Rank of block $B \geq T_R$ and $C_B \leq T_C$ **then**
    $\{a, ab_1, ..., ab_6, b\} \leftarrow \text{LI}(a, b, 6)$
    $\{ax_2, abx_{21}, ..., abx_{24}, bx_2\} \leftarrow \text{LI}(ax_2, bx_2, 4)$
    $\{ax_1, abx_{11}, abx_{12}, bx_1\} \leftarrow \text{LI}(ax_1, bx_1, 2)$
  else if Rank of block $B \geq T_R$ and $C_B > T_C$ **then**
    $\{a, ab_1, ab_2, ab_3, B_{51}\} \leftarrow \text{LI}(a, B_{51}, 3)$
    $\{ax_2, abx_{21}, abx_{22}, B_{61}\} \leftarrow \text{LI}(ax_2, B_{61}, 2)$
    $\{ax_1, abx_{11}, B_{71}\} \leftarrow \text{LI}(ax_1, B_{71}, 1)$
  end if
end if
Run Steps 10-18 to compute the elements with suffix $ac-$
Compute the elements of subblocks B, C and D with the logic as in Steps 9-20

Terminal elements in boundary blocks such as block A
if Rank of block $A \geq T_R$ and $C_A \leq T_C$ **then**
  Non-corner elements A_{ij} (i=1:4, j=5:8; i=5:8, j=1:4)
  for $i = 1$ to 4 **do**
    $\{A_{i5}, ..., A_{i8}\} \leftarrow \{a, ab_1, ab_2, ab_3\}$
    $\{A_{5i}, ..., A_{8i}\} \leftarrow \{a, ac_1, ac_2, ac_3\}$
  end for
  Corner elements A_{ij} (i=1:4, j=1:4)
  for $i = 1$ to 4 **do**
    $\{A_{11}, A_{22}, ..., A_{55}\} \leftarrow \text{LI}(A_{11}, A_{55}, 3)$
    Compute the other non-corner elements with the logic as in Steps 15-17
  end for
end if

example, Rank 5. They contribute an insignificant increase in bits (much less than 1 Kbits) and their DCT coefficients are all discarded. This illustrates that the proposed scheme can encode regions-of-interest with arbitrary shapes.



33.36 | 0.732 | 0.19 33.43 | 0.769

32.21 | 0.740 | 0.20 32.28 | 0.805

Fig. 13: Some examples of the worst distortion (left) and the improved quality after de-pixelization (right) for FRUIT (top) and PEPPER (bottom) images. The numbers denote PSNR and MSSIM, respectively.

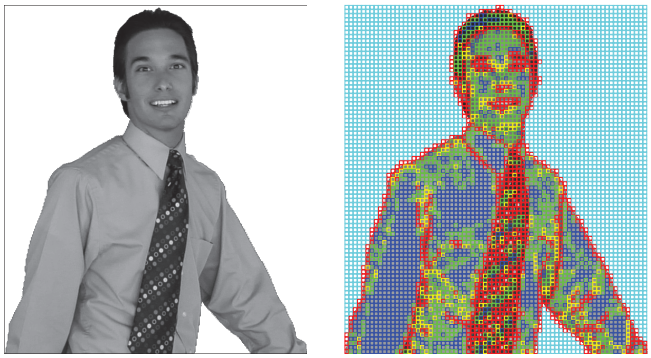


Fig. 14: An image with a segmented object as part of a video frame from a CS's surface (left). The blocks after applying the proposed block ranking algorithm (right). Image border is added for better view by readers.

IV. Algorithm Complexity and FPGA Design

We have proposed an FPGA-based platform for the design and implementation of a DMP network node [31] (Fig. 15). It provides a detailed introduction to the platform architecture and the simulation-implementation environment for the design. Our compact implementation on a Xilinx Virtex-6 ML605 board consumes very small amount of the available resources. Moreover, the elementary operations in our implementation take (much) less than 5 μs as desired to meet the low-latency

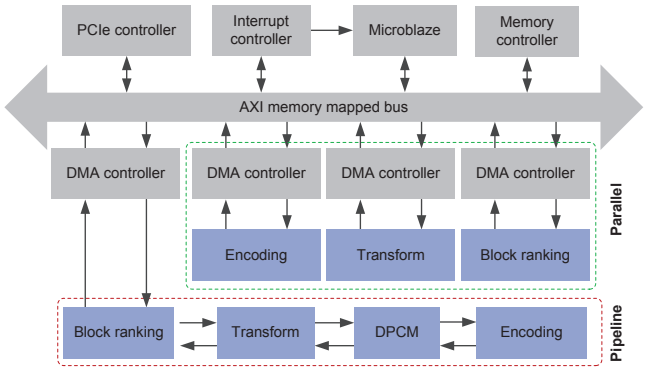


Fig. 15: The FPGA-based architecture of a DMP transmitter with the pipeline and parallel approaches.

requirement. The AXI bus and the EDK environment are used to implement both the transmitter and receiver in DMP. Although the architecture of the access node has different number of compression-scheme components than that of the network node, their core components and adopted processing approach are the same. In addition to controlling the data flow within the FPGA system, Microblaze is also used to establish and maintain the communication with the external DMP servers located on a host (PC machine).

The design's modularity and scalability ease the integration of the external modules into the platform, which can follow parallel, pipeline or hybrid approaches. The first two approaches are depicted in Fig. 15. By assuming equal access in the memory-mapped AXI bus, the parallel approach offers flexibility because it permits software elimination in certain steps of the processing chain if necessary, e.g. the encoding. This is possible because the Microblaze governs all the execution steps of the chain, and they are independently connected to a single AXI bus. On the other hand, the pipeline implementation is more efficient provided that all the modules are used in the processing chain and the pipeline latency is not critical. Adopting both approaches in a hybrid fashion is also an alternative depending on the application.

A. Calculation of Entropy

The complexity for HW design of the major parts of the proposed system is presented as follows. Fig. 16 (a) shows the entropy module, and the dashed line covers the parallel structure. The logarithm operation can be implemented as a registered LUT for 8-bit input data at one clock (CLK) [2]. Thus, the overall latency is 7 CLK, i.e. 3 CLK for each multiplier, and given n parallel structures, it becomes $1 + 3\log(n)$ clocks. Consisting of a block RAM (BRAM) memory and an incremental logic, a histogram module with 64 input integer values provides the probability values p_i.

Fig. 16 (b) and (c) show the simplified and real diagrams, respectively. The input data for the evaluation of the histogram address the BRAM and the BRAM's D_{out} stores the count of $Data_{\text{in}}$'s prior to the occurrences at the BRAM address bus. The counter is incremented by one and written back to the BRAM at the same address. The BRAM limits the calculation speed as the output data is one CLK delayed with respect to

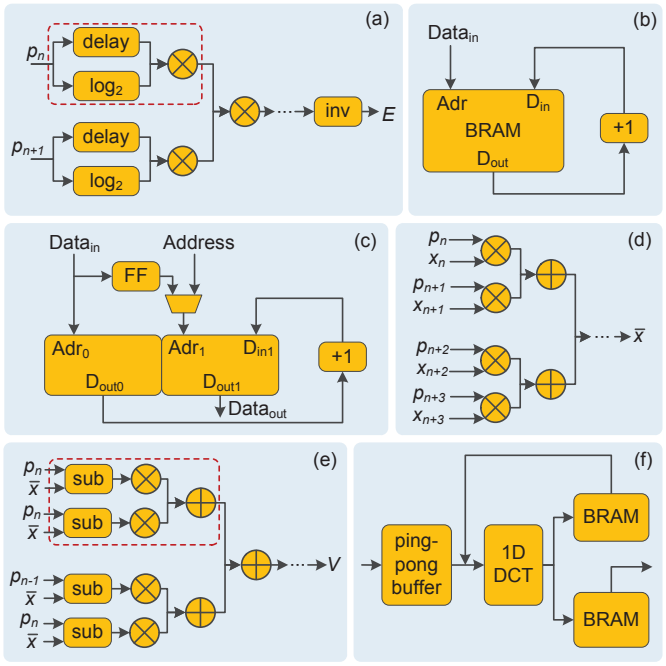


Fig. 16: The modules for calculating entropy (a), histogram (b,c), mean(d), variance (e), and 2D-DCT (f).

the address bus (a synchronous memory data read); hence, evaluating a single input pixel involves two CLK. Strong parallelization of the computations is possible [12], and the histogram computation needs 34 LUTs and 23 FFs.

B. Calculation of Mean and Variance

Computing mean values of n inputs of $p_i x_i$ (Fig. 16 (d)) takes 3+log(n) CLK, and the variance-calculation module consists of the mean-calculation unit and a set of subtractors, multipliers and adders. They are strongly parallel modules which process the data every clock cycle. The parallelization determines the computation time, and generally it is 4+log(n) CLK plus the latency from the mean-calculation module.

C. Calculation of 2D-DCT, IDCT and DPCM

By employing a two-pass 1D-DCT transform [28], computing a complete 8×8 2D-DCT takes 80 clock cycles and can work at 107 MHz. Adopting a ping-pong fashion, it stores the results of the 1D-DCT by means of an intermediate buffer (Fig. 16 (f)). It is a trade-off between resource consumption and speed which complies well with the idea of an AXI-based Microblaze-controlled architecture. Nevertheless, other implementation approaches for 2D-DCT can be considered, such as replacing the time-consuming multiplications with LUT accesses [14]. As for the DPCM, its sequential execution flow favors software implementation in Microblaze, and the processing power will not be absorbed because DC coefficients are fewer than AC coefficients.

D. Encoding and Decoding

Encoding Fibonacci code is simple, but straightforward implementation in iterative procedures needs substantial clock

cycles. Therefore, it is better implemented as LUT and executed in one clock, which is feasible because Fibonacci coder for 8-bit numbers consumes merely 8×12 bits, and 3072 bits can fit into a single BRAM memory of 18 Kbits. Thus, it occupies only 2 BRAMs for both encoding and decoding. Moreover, the *bzip2* algorithm can be implemented in software in Microblaze.

E. Packet Dropping

The dropping module in Fig. 17 is the core component of the QS scheme in a DMP node. The module is integrated to the platform in Fig. 15 also via a direct memory access (DMA) controller. Our strategy is to extensively use AXIS (AXI Streaming) bus which provides system flexibility. All the modules connected to the network node are AXIS-compatible.

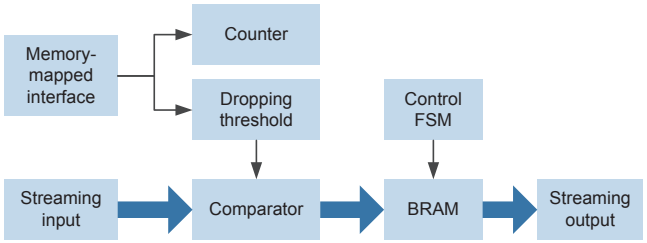


Fig. 17: The proposed structure for DMP dropping module.

The dropping module works as follows. The network packets carrying image data are sent over the PCIe to the external memory (DDR3 in ML605) and stored on a long queue. The Microblaze monitors the status of the queue, programs DMA controller to read the data from the external memory, and writes them to the dropping module. Based on the data received from the other nodes in the network, a current threshold value for dropping is computed and written to the internal register of the dropping module. The data fed into the dropping module from the external memory by the DMA are either dropped or passed through to the internal memory (BRAM) after compared with the threshold. Once the DMA write-operation is finished, the Microblaze is interrupted and informed that the dropping statistics can be read from the internal register of the dropping module. The Microblaze then programs the DMA controller based on the statistics, and the data stored in the internal memory are transferred to the external memory. The dropping process is finished when the data is read from the internal memory of the dropping module and written to the external RAM. All HW modules in both access and network nodes are interconnected with AXIS bus and controlled by the Microblaze.

F. Depixelization as Post-Processing

Depixelization is essentially a finite-impulse response (FIR) filter operation conducted on block borders. Xilinx delivers a FIR compiler tool which can be used to compile the FIR architectures to generate the depixelization filter [38]. It includes 12-bit coefficients and 60-tap input data which can work at 150 MHz and consume 3382 LUT-FF pairs.

G. Overall Performance

The system performance strictly depends on the chosen architecture (parallel, pipeline or hybrid) and the synchronization between the modules. The following is the gross estimate of the computational time for processing a video frame as a 1920×1080 color image. By assuming 256-bit AXI bus width, 64-pixel block, and 100 MHz FGPA clock cycle as the main constraints, the internal processing speed per single thread becomes 50 Mblocks/s. As the luma and 25% subsampled chroma images equal 48,600 blocks, the essential processing time becomes 1 ms per frame. Since the object-based processing reduces the number of blocks processed per frame, the processing time per frame is less than 1 ms.

The consumed resources are detailed in Table II, and the additional pre- and post-processing steps are excluded. LUTs consume the most resources which are roughly 10% of all the XCVLX240T resources, the FPGA used in ML605 [36]. It is possible to balance the usage of LUTs with DSP implementation to equalize resource consumption. Consequently, 15 to 20 parallel processing streams can be implemented as in Fig. 15 which reduce the processing time to approximately 50 μs per frame. Moreover, several FPGA boards can be used as a one-stop system [16] to achieve chip-level parallelism.

TABLE II: Total consumption of resources

Module	#LUT	#FF	#BRAM
Entropy	$m{\cdot}\log(m){\cdot}115 + n \cdot 19$	$110 + n \cdot 64$	0
Histogram	34	23	1
Variance	$n \cdot 115 + [n + n{\cdot}\log(n)] \cdot 8$	$n \cdot 110$	0
DCT	123	110	2
Fibonacci	0	0	2

m denotes the number of parallel inputs for entropy module

V. Conclusion and Future Work

We have presented an ultrafast, DCT-based, embedded image-compression scheme which is quality scalable and can process objects with arbitrary shapes. It is designed for network architecture such as DMP that guarantees maximum EED. The encoder mainly consists of block ranking, 2D-DCT and entropy coding. As the quantization as in existing image coding schemes is not present, the main loss of information in this approach is due to the intelligent dropping of data packets by network nodes during transmission to guarantee local delay. Since simplicity and parallelization are favored for minimizing processing time, block entropy and variance are used for block ranking, which work satisfactorily by yielding four ranks of 8×8-blocks with increasing importance. Universal codes are employed to encode the resulting block ranks and DCT coefficients. The VQ in PSNR and MSSIM of several common test images due to dropping is given against the bitrates, which are also compared to the results from JPEG. JPEG performs better as expected because it can exploit global redundancies in an image and the bitstream, but lacks the scalability. Fundamental differences between the two schemes make such a performance comparison essentially irrelevant. Excessive dropping results in pixelation artifacts that are faithfully contained in the blocks which the receiver can immediately locate from the side information available in the

packet headers of the remaining bitstream. A depixelization algorithm, a post-processing step at the receiver, is proposed for the worst distortion. We show how the scheme can be applied to objects of arbitrary non-rectangular areas in images after segmentation. Every video frame, channel, segmented object, and block are processed independently, allowing fully parallel HW implementation. Finally, as indicated by the estimated complexity and resource consumption of the proposed scheme for FPGA implementation, a video frame as a 1920×1080 color image can be processed, encoded and decoded in less than 1ms, sufficient to meet the maximum EED at 11.5ms. Ideas for further performance improvement include incorporating fast intra-prediction and advanced depixelization.

REFERENCES

[1] N. Ahmed, T. Natarajan, R. K. Rao, "Discrete cosine transform," *IEEE Trans. Computers*, vol. 23, no. 1, pp. 90–93, 1974.

[2] N. Alachiotis, A. Stamatakis, "Efficient floating-point logarithm unit for FPGAs," in *Proc. IEEE Int'l Sym. Parallel & Distributed Processing, Workshops and PhD Forum (IPDPSW)*, 2010.

[3] M. Burrows, D. Wheeler, *A Block Sorting Lossless Data Compression Algorithm*, Technical Report 124, Digital Equipment Corporation, 1994.

[4] C. Chafe, M. Gurevich, G. Leslie, S. Tyan, "Effect of time delay on ensemble accuracy," in *Proc. Int'l Symp. Music Acoustics*, 2004.

[5] T.A. DeFanti, D. Acevedo, R.A. Ainsworth, M.D. Brown, S. Cutchin, G. Dawe, K.U. Doerr, A. Johnson, C. Knox, R. Kooima, F. Kuester, J. Leigh, L. Long, P. Otto, V. Petrovic, K. Ponto, A. Prudhomme, R. Rao, L. Renambot, D.J. Sandin, J.P. Schulze, L. Smarr, M. Srinivasan, P. Weber, G. Wickham, "The future of the CAVE," *Central European J. Eng.*, vol. 1, no. 1, pp. 16–37, 2011.

[6] Y. Du, J. Wang, S.-M. Guo, P.D. Thouin, "Survey and comparative analysis of entropy and relative entropy thresholding techniques," *IEE Proc. - Vision, Image and Signal Processing*, vol. 153, no. 6, pp. 837–850, 2006.

[7] P. Fenwick, Universal Codes, in K. Sayood (ed.), *Lossless Compression Handbook*, Academic Press, 2003.

[8] A.S. Fraenkel, S.T. Klein, "Robust universal complete codes for transmission and compression," *Discrete Applied Mathematics*, vol. 64, no. 1, pp. 31–55, 1996.

[9] S. W. Golomb, "Run-length encodings," *IEEE Trans. Information Theory*, vol. 12, no. 3, pp. 399–400, 1966.

[10] P. Holub, J. Matela, M. Pulec, M. Šrom, "Ultragrid: low-latency high-quality video transmissions on commodity hardware," in *Proc. ACM Multimedia*, 2012, pp. 1457–1460.

[11] *Advanced Video Coding for Generic Audio-Visual Services*, ITU-T Rec. H.264 and ISO/IEC 14496-10 (AVC), ITU-T and ISO/IEC JTC 1, May 2003 (and subsequent editions).

[12] E. Jamro, M. Wielgosz, K. Wiatr, "FPGA implementaton of strongly parallel histogram equalization," in *Proc. IEEE Design and Diagnostics of Electronic Circuits and Systems (DDECS)*, 2007, pp. 1–6.

[13] J. Kopf, D. Lischinski, "Depixelizing pixel art," *ACM Trans. Graphics*, vol.30, no. 4, pp. 99:1–99:8, 2011.

[14] R. Kutka, "Fast computation of DCT by statistic adapted look-up tables," in *Proc. IEEE Int'l Conf. Multimedia and Expo (ICME)*, 2002, pp. 781–784.

[15] J-R. Ohm, "Advances in scalable video coding," *Proc. IEEE*, vol. 93, vol. 1, pp. 42–56, 2005.

[16] One Stop Systems, http://www.onestopsystems.com/

[17] L. A. Rønningen, *The Distributed Multimedia Plays Architecture (version 3.20)*, Technical Report, ITEM, NTNU, 2011.

[18] L.A. Rønningen, O.J. Wittner, "Experiments on remote conducting between Trondheim and Lisbon," ITEM, NTNU, 2011.

[19] H. Schwarz, D. Marpe, T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, 2007.

[20] J. Seward, bzip2 codec, http://www.bzip.org/

[21] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[22] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3445–3462, 1993.

[23] D. Solomon, G. Motta, *Handbook of Data Compression*, Springer, 2010.

[24] G. Sorwar, A. Abraham, L.S. Dooley, "Texture classification based on DCT and soft computing," in *Proc. 10th IEEE Int'l Conf. Fuzzy Systems*, 2001.

[25] H. Sun, A. Vetro, J. Xin, "An overview of scalable video streaming," *Wireless Communications and Mobile Computing*, vol. 7, pp. 159–172, 2007.

[26] TGFX. http://www.timelinegfx.com/.

[27] R. Tucker, "The role of optics and electronics in high-capacity routers," *J. Lightwave Tech.*, vol. 24, no. 12, pp. 4655–4673, 2006.

[28] A. Tumeo, M. Monchiero, G. Palermo, F. Ferrandi, D. Sciuto, "A pipelined fast 2D-DCT accelerator for FPGA-based SoCs," in *Proc. IEEE Computer Society Annual Symp. VLSI*, 2007, pp. 331–336.

[29] S. Vajda, *Fibonacci and Lucas Numbers, and the Golden Section Theory and Applications*, Ellis Horwood, Chichester, 1989.

[30] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, pp. 600–612, 2004.

[31] M. Wielgosz, M. Panggabean, J. Wang, L. A. Rønningen, "An FPGA-based platform for a network architecture with delay guarantee," *J. Circuits, Systems and Computers*, vol. 22, no. 06, 2013.

[32] P. Schelkens, A. Skodras, T. Ebrahimi, *The JPEG 2000 Suite*, Wiley, Series: Wiley-IS&T Series in Imaging Science and Technology 2009.

[33] D. Taubman, "High performance scalable image compression with EBCOT," *IEEE Trans. Image Processing*, vol. 9, no. 7, pp. 1158–1170, 2000.

[34] D. Taubman, M. Marcellin, *JPEG 2000: Image Compression Fundamentals, Standards and Practice*, Kluwer Academic Publishers, 2001.

[35] R.J. Van der Vleuten, R.P. Kleihorst, C. Hentschel, "Low-compexity scalable DCT image compression," in *Proc. IEEE Int'l Conf. Image Processing*, 2000, pp. 837–840.

[36] Virtex series, http://www.xilinx.com/publications/matrix/Virtex_Series.pdf

[37] D. Wu, Y. Hou, Y-Q. Zhang, "Transporting real-time video over the Internet: challenges and approaches," *Proc. IEEE*, vol. 88, no. 12, pp. 1855–1875, 2000.

[38] Xilinx, http://www.xilinx.com/support/documentation/ip_documentation/fir_compiler_ds534.pdf

[39] x264, http://www.videolan.org/developers/x264.html

[40] Z. Yang, B. Yu, W. Wu, K. Nahrstedt, R. Diankov, R. Bajscy, "A study of collaborative dancing in tele-immersive environments," in *Proc. 8th IEEE Int'l Symp. Multimedia*, 2006, pp. 177–184.

[41] J. Zhang, T. Tan, "Brief review of invariant texture analysis methods," *Pattern Recognition*, vol. 35, no. 3, pp. 735–747, 2002.