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

Volume 2 Issue 11

November 2011

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

SAI

www.ijacsa.thesai.org

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS

A Publication of

The Science and Information Organization

IJACSA Editorial

From the Desk of Managing Editor…

It is a pleasure to present our readers with the November 2011 Issue of International Journal of Advanced Computer Science and Applications (IJACSA).

The renaissance stimulated by the field of Computer Science is generating multiple formats and channels of communication and creativity. IJACSA is one of the most prominent publications in the field and engaging the ubiquitous spread of 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.

The journal has a wide scope ranging from the many facets of methodological foundations to the details of technical issues and the aspects of industrial practice. It includes articles related to research findings, technical evaluations, and reviews. In addition it provides a forum for the exchange of information on all aspects.

The editorial board of the IJACSA consists of individuals who are committed to the search for high-quality research suitable for publication. These individuals, working with the editor to achieve IJACSA objectives, assess the quality, relevance, and readability of individual articles.

The contents include original research and innovative applications from all parts of the world. This interdisciplinary journal has brought together researchers from academia and industry as well as practitioners to share ideas, problems and solutions relating to computer science and application with its convergence strategies, and to disseminate the most innovative research. As a consequence only 26% of the received articles have been finally accepted for publication.

Therefore, IJACSA in general, could serve as a reliable resource for everybody loosely or tightly attached to this field of science.

The published papers are expected to present results of significant value to solve the various problems with application services and other problems which are within the scope of IJACSA. In addition, we expect they will trigger further related research and technological improvements relevant to our future lives.

We hope to continue exploring the always diverse and often astonishing fields in Advanced Computer Science and Applications.

Thank You for Sharing Wisdom!

Managing Editor
IJACSA
Volume 2 Issue 11, November 2011
editorijacsa@thesai.org
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

IJACSA Reviewer Board

CONTENTS

Crytosystem for Computer security using Iris patterns and Hetro correlators

R. Bremananth
Information Systems and Technology Department,
Sur University College,
Sur, Oman.

Ahmad Sharieh
Information Systems & Technology Department,
Sur University College,
Sur, Oman.

***Abstract*—Biometric based cryptography system provides an efficient and secure data transmission as compare to the traditional encryption system. However, it is a computationally challenge task to solve the issues to incorporate biometric and cryptography. In connection with our previous works, this paper reveals a robust cryptosystem using iris biometric pattern as a crypto-key to resolve the issues in the encryption. An error correction engine based on hetro-correlators has been used to evoke the partially tarnished data fashioned by the decryption process. This process determines the non-repudiation and key management problems. The experimental results show that the suggestion algorithm can implement in the real-life cryptosystem.**

Keywords-Auto-correlators; Biometric; crytosystem; Hetro-correlators.

I. INTRODUCTION

Cryptography provides a secure proliferation of information exchange across the insecure data communication [1]. It authenticates messages based on the mathematical key but not based on the real-life user those who are the genuine owner. Traditional cryptosystem requires a lengthy key to encrypt and decrypt in sending and receiving the messages, respectively. But these keys can be guessed or cracked. Moreover, maintaining and sharing lengthy, random keys in enciphering and deciphering process is the critical problem in the cryptography system. A new approach is described for generating a crypto key, which is acquired from iris patterns. In the biometric field, template created by the biometric algorithm can only be authenticated with the same person. Among the biometric templates, iris features can efficiently be distinguished with individuals and produces less false positives in a large population. This type of iris code distribution provides merely less intra-class variability that aids the cryptosystem to confidently decrypt messages with an exact matching of iris pattern. In traditional cryptography system, key management is a cumbersome process that is, key must be generated each time with an extensive computational process and the dissemination of keys is also a very difficult process at the non-secure channels [1]. It consumes lot of system time and produces overburden to the application domains. In addition, non-repudiation cannot easily be handled in the traditional cryptosystem.

The Biometric key cryptography (BKC) is an emerging reliable alterative that can be used to resolve key management, large key computational process and address the non-repudiation problems [2]. In the cryptography system, data will be secured using a symmetric cipher system and in public-key system digital signatures are used for secure key exchange between users. However, in both systems the dimension of security accuracy is dependent on the cryptography strong keys. They are required to remember and enter the large key whenever needed. Instead of remembering large keys, the user may opt to give password to encrypt and decrypt the cryptography keys. There is no direct tie up between user and password that is, the system running the cryptography algorithm is unable to differentiate the genuine user and impostors who are unauthorized to work with the system.

Thus, a reliable alternative to the password security is the biometric guard for the cryptography keys. Whenever user wishes to access through a secured key, biometric sample is captured, authenticated by the classifiers and then key is released to encipher / decipher the desired data. In general biometric cryptosystem has been classified by three categories. The first method is to release the cryptography key from secure area in accordance with biometric matching algorithm. It requires the secured communication line to avoid eavesdropper's attacks. Furthermore, if the user may store the biometric templates or crypto keys in workstation machines then the system becomes an insecure one. In the next method, the crypto key is embedded as a part of biometric template in a specific location. However, if impostors may determine the location of the keys, again it becomes catastrophic to the system. The third method is based on using biometric features as cryptography keys, which gives more secure manner of proliferation of information exchange.

The proposed approach is broadly classified into three phases. The first phase is related with compact way to obtain iris feature codes from the human irises. The second one describes the algorithm to encrypt and decrypt the messages using iris bits. In the third phase, the error correction engine is employed to recall the partially corrupted bits generated in the decryption using associative memories. The issue of biometric pattern is the partially varied features produced in the feature extraction process, which subsequently makes partially corrupted data in the decryption process. This dissimilarity may occur due to environments, illuminations, distance variation and other artifacts. However more stable pattern produced by the iris is secured in the person's lifetime and produces limited

number of bits variations in the features, which assists to decrypt the messages in massive manner. In addition, re-enrolment of iris keys is required to preserve the system security more consistently.

In the current literature several studies were proposed related with biometric cryptosystem but most of them dealt with fingerprints and few of them were concerned with iris features. Albert Bodo proposed a method of directly using biometric as cryptography key in the patent of German [1]. In (Davida et al. [2][3]), 2048-bit iris code was used for enciphering and deciphering process. Key generation is invoked based on the error bits of the iris codes. This system stored the error correction bits along with iris keys inside the database. Thus, impostors may eavesdrop key information and a count of error correction bits from the local database. In (Linnartz et al. [4], Clancy et al. [5], Monrose et al. [6]), the key generation was based on biometrics such as fingerprints [18] and voices, but they required more calculations to release the key than the traditional cryptography system. The problem of generating cryptograph key from face biometric features had been studied by Yao-Jen Chang et al. [7]. The survey of multi-biometric cryptosystems was discussed by Uludag et al. [8]. A method of iris compression for cryptography documentation on off-line verification was proposed by Daniel et al. [9]. In this study, a modified Fourier-Mellin transformation was employed to create iris template for representing EyeCert system, which consists of two components. The first one is details of personal data related with the subjects, and the second one is the iris feature encoded in the form of barcodes. In another study of iris biometric cryptosystem, Feng Hao et al. [10] proposed a method based on error-free iris key that was devised using a two-layer error correction technique incorporated with Hadamard and Reed-Solomon codes. The extracted code was saved in a tamper-resistant token such as a smart card. In our previous work, (Bremananth et al. [11]) proposed auto-correlator to recoup the corrupted bio-metric crypto key. In this paper, a robust hetro-correlator has been proposed to regain the data.

The block diagram of the proposed iris cryptosystem is illustrated in Fig. 1. It suggests a compact way to extract feature from the iris patterns and these features are treated as crypto key for the on-line cryptography system. This system outperforms other traditional approaches and provides an efficient solution for non-repudiation approach as well. It employs 135-bit iris code which is extracted by wavelet analysis[12][13][14] and applying these codes in enciphering and deciphering of the input stream of binary data which might be originating from voice, text, video, image or other sources. Next, the auto-correlators and hetero-correlators are used to recall original bits from the partially corrupted data produced in the decryption process. It intends to resolve the repudiation and key management problems. However, the performance of error correction model dependents on the correlators used in the system. Hence the guarantee issues of these methods were verified and the experimental results were analyzed in both symmetric iris cryptosystem (SIC) and non-repudiation iris cryptosystem (NRIC). It shows that this new approach provides considerably high authentication in enciphering and deciphering processes. The remainder of the paper has been organized as follows. Section II describes the symmetric iris cryptosystem. Non-repudiation cryptosystem is described in Section III. Error correction engines and their functionalities are given in Section IV. Section V describes the experimental results of the bio-metric cryptosystem and concluding remarks are given in Section VI.

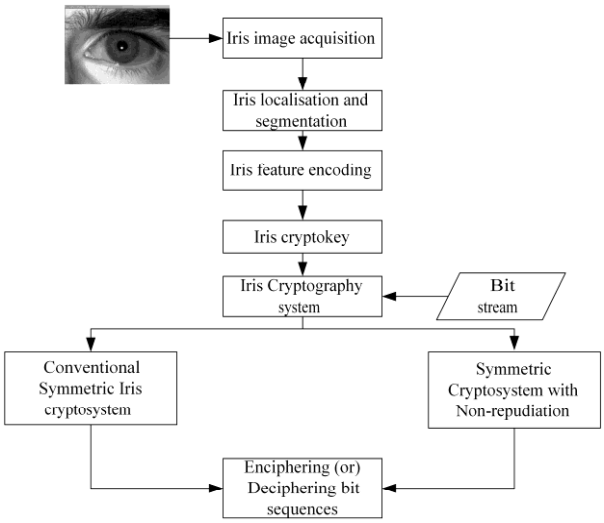


Figure 1. A proposed block diagram of the iris cryptography system.

II. SYMMETRIC IRIS CRYPTOSYSTEM

Iris patterns are used for fabricating a key to encipher and decipher the plain text in between sender and receiver over insecure channels [2][11]. The advantages of iris cryptosystem are to reduce the system processing time to make a complex key for standard cryptography algorithm and to generate cipher keys without getting back from complex key generation sequences. The identical iris code is used in both ends to encrypt and decrypt the message in the SIC system. In order to decrypt a message, the recipient needs an identical copy of the iris code. Figure 2 shows the iris based symmetric cryptography system. The transmission of enrolled iris code over the channel is vulnerable to eavesdropping. Hence, the copy of the enrolled iris code is needed in the recipient side, which is being used by the decryption process. In this approach, XOR operation is used to encrypt and decrypt the message. The significant steps of SIC encryption algorithm is described as follows:

Step 1: Let K be the key sequence $I_1, I_2, ..., I_p$ produced by iris feature encoding algorithm for the encryption transformation. In the experiment 136-bit key sequence (135-bit iris code and one padding bit) is used in the encryption process.

Step 2: Let S be a source alphabet of N symbols $S_1, S_2, ..., S_N$. Each alphabet in S is converted to its equivalent 8-bit binary strings. The bits of messages undergo XORing with iris key sequence and generate a non-breakable cipher-bit described as

$$C_i = Ency(S_1, S_2, ..., S_N \oplus I_1, I_2, ..., I_p) \qquad (1)$$

where C_i is set of cipher bits. The decryption algorithm is described as follows:

Step 1: The testing iris pattern is extracted and iris codes are formed. The iris-matching algorithm verifies the test and the enrol iris codes. If weighted distance (WD) is $0 \le \mathrm{WD} \le 0.19$, then the matched enrolled iris code is used for deciphering the messages, otherwise rejected.

Step 2: Let $I_1, I_2, ..., I_p \in K$ be an enrolled iris code and $C_1, C_2, ..., C_n$ is a set of cipher text produced by the encryption process. Enrolled iris codes are XORed with set of cipher bits and generate the original messages using Equation (2).

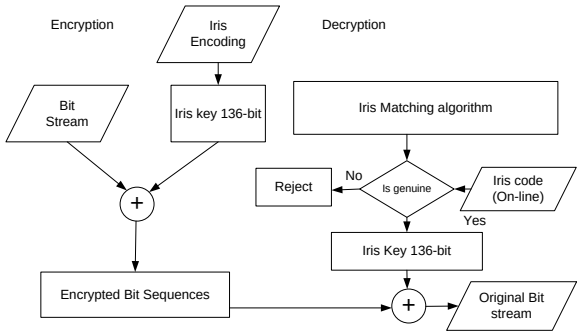


Figure 2. The process of SIC system.

$$S_i = Decy(C_1, C_2, ..., C_n \oplus I_1, I_2, ..., I_p) \quad (2)$$

where S_i is set of source alphabet bits and I = 1,2, 3, … N. In the SIC system, key dissemination problem is completely avoided. However, the system needs iris database and iris-matching algorithm in the decryption process to get back the original messages. In order to resolve repudiation problem, the iris database and iris-matching algorithm are eliminated from the SIC system. The detailed description of this process is discussed in the next section.

III. Non-Repudiation Iris Cryptosystem

Unlike SIC system, the NRIC system bypasses the iris-matching process and do not access iris database in the decryption process. The testing iris code can directly be XORed with cipher bits transmitted by the encryption process as illustrated in Fig. 3. Iris codes are changed from session to session with minimum variation (WD<=0.19) for the same subject eye. Hence the decryption process may produce the probability of partially corrupted cipher bits ranging from 0 to 0.19. Perhaps, if intruder may tap the cipher bits at the non-secure channels then the probability of decrypting the message is complicated from 0.2 to 1 partially corrupted bit in every 135-bit iris code. Thus, it produces more complexity to the intruder to get back the original messages. But the cipher bits accessed by the genuine subjects have probability of error rate at most 0.19, so that, less complexity have been created in the decryption process.

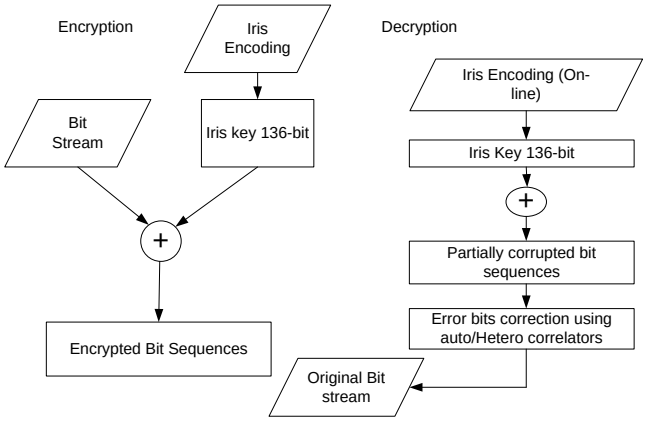


Figure 3. A sequence sketch of Non-repudiation cryptosystem.

In this method cipher bits are directly XORed with the test iris key and produce the partially corrupted bits. These are very close to the original message if the test iris key is actually extracted from the genuine subject; otherwise the partially corrupted bits are larger than the threshold maintained in the system.

Thus impostors can be restricted to access the original scripts. The error bit correction module subsequently corrects these bits by using the two different correction engines such as either auto-correlators or hetero-correlators that perform the probability of error correction based on iris-weighted distance. Thus this process overcomes repudiation problem and reduces the key management issues. However, the performance of the NRIC fully depends on the guarantee of the error correction engines because recalling the original bits is a difficult process in the real time processing of encryption and decryption.

IV. Error Correction Engines

In the process of biometric cryptosystem, the major limitation is a way to get back the original bits from the partially corrupted bits generated by the decryption. In the literature, several studies had been performed to recall the trained patterns from the partially corrupted patterns. Bart Kosko et al. [15] enhanced the bidirectional associative memories (BAM), which behaves as a hetero-associative content addressable memory (CAM) storing and recalling the vector pairs.

The bidirectional associative memory with multiple training can be guaranteed to recall a single trained pair under suitable initial conditions of data. Sufficient condition for a correlation matrix to make the energies of the training pairs was described by Yeou-Fang et al. [16]. An essential condition for generalization of correlation matrix of BAM which guarantees the recall of all the training pairs was discussed by Yeou-Fang et al. [17]. This paper adopts two different methods to recall the corrupted patterns. The first one is related to auto-associative and the other one is concerned with hetero-associative.

A. Autocorrelators

Associative memories are one of the key models of neural network and they can act as a human brain to recall the associated patterns perfectly from the corrupted patterns. If the associated pair (x, y) is the identical pattern, then the model of associative memory is called as auto-associative memory. For the recall operation, auto-associatives require the correlation memory or connection matrix, which aids to retrieve original patterns from the partially corrupted pattern. It is called as auto-correlators and is adopted in the error correction process of NRIC. The algorithm of error bits correction process is described as follows [11]:

Step 1: The partially corrupted data obtained in the decryption process is taken for further processing. This data is transformed to bipolar patterns (ϕ_c). Let M be the number of stored bipolar patterns $p_1, p_2, \ldots, p_m$ and i[th] patterns is $(p_{i1}, p_{i2}, \Lambda, p_{in})$ where n is the number of bits in the stored pattern. The connection matrix CM is derived as

$$CM_{ij} = \sum_{i=1}^{n} \left[p_i^T \right] \left[p_i \right] \text{ for } i = 1..n, \text{ for } j = 1..n \quad (3)$$

Step 2: The auto-correlator recalls the original patterns (θ) using

$$\theta_j = g((\phi_{cj} * CM), p_j) \quad for\ j = 1..m \quad (4)$$

$$g(\chi, \varphi) = \begin{cases} 1 \ if \ \ \chi > 0 \\ \varphi \ if \ \ \chi = 0 \\ -1 \ if \ \ \chi < 0 \end{cases} \quad (5)$$

where θ_j is the recalled original pattern, ϕ_c is a partially corrupted data and $g(\chi, \varphi)$ is the threshold function.

Step 3: Repeat Step 2 until $\sum_{i=1}^{n} |\phi_i - \theta_i| > \rho$, where ρ is a vigilance parameter.

The parameter ρ provides minimum error bit correction in between the genuine subject iris code and partially corrupted cipher bits. This parameter gives more complexity to the intruder to get back the original messages. For example, if the patterns are $p_1 = [1 \quad 1 \quad -1]$, $p_2 = [-1 \quad -1 \quad 1]$, $p_3 = [1 \quad -1 \quad 1]$ then the connection matrix (CM) is: $\begin{bmatrix} 3 & 1 & -1 \\ 1 & 3 & -3 \\ -1 & -3 & 3 \end{bmatrix}$

If partially corrupted data produced in the decryption process is $p = [-1 \quad 1 \quad 1]$ then the computation with CM produce the threshold conditions: g(-3,-1),g(-1,1) and g(1,1). It gives the original pattern $O = [-1 \quad -1 \quad 1]$.

B. Heterocorrelators

In this approach, noisy variation of different types of iris codes are not explicitly estimated and stored in the verification database [17]. If they may explicitly be estimated, then it leads to leak of security information to the adversary. Hence, hetero-correlations are directly used to recall the original patterns from the corrupted patterns that need not have any additional information such as noisy variations. This is nothing but an associative memory, which is an imitation model of human brain's ability to recall associate patterns. In the non-repudiation cryptosystem, the decryption produces noise bits which should be corrected properly and converted to its real bit sequences. If the associated pattern pairs (x, y) are different, then this model recalls y. If x is given, then y can be called. This is referred as hetero-associative memory. This memory is used to recall the original patterns from the corrupted patterns. For the recall operation, hetero-associative requires a correlation memory or connection matrix, which aids to retrieve original patterns. This is so-called hetero-correlators. The algorithm of error bits correction process is described as follows:

Step 1: The partially corrupted data obtained in the decryption process is taken for further processing. This data is transformed into its bipolar patterns (δ). Let M be the number of stored bipolar pairs given as

$$\left\langle \{P_1, Q_1\}, \{P_2, Q_2\}, ..., \{P_m, Q_m\} \right\rangle \quad (6)$$

where $P_i = \{p_{i1}, p_{i2}, \ldots p_{in}\}, Q_i = \{q_{i1}, q_{i2}, .. q_{io}\}$, P and Q represent stored and exemplar patterns of distorted bipolar data, respectively. The connection matrix (CM) is derived as

$$CM_{ij} = \sum_{i=1}^{n} E_i \left[P_i^T \right] \left[Q_i \right] \text{ for } i = 1...n, \text{ for } j = 1..o \quad (7)$$

where CM is a correction matrix used in the hetero-correlation process and E is a set of energy constants i.e., $E \in R^+$, R is a set of real numbers. Calculate κ' and κ from Equations (8) and (9) and assign to δ' and δ, respectively.

Step 2: The hetero-correlator recalls the original bit sequences (φ) using

$$\kappa' = \Theta(\delta \bullet CM) \tag{8}$$

$$\kappa = \Theta(\kappa' \bullet CM^T) \tag{9}$$

$$\Theta(\lambda) = \varphi = \varphi_1, \varphi_2, \ldots, \varphi_n \tag{10}$$

$$\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\} \tag{11}$$

$$\varphi_i = \begin{cases} 1 \ if \ \ \lambda_i > 0 \\ \varphi_i \ if \ \ \lambda_i = 0 \\ -1 \ if \ \ \lambda_i < 0 \end{cases} \tag{12}$$

where δ is a set of partially corrupted bipolar bits generated by the decryption process, Θ is a threshold function of hetero-correlation, λ represents multiplication result of the correction matrix for the given distortion bit patterns, φ is set of the recalled bits, κ' represents result of exemplars and κ is a sequence of corrected bits.

Step 3: After performing error correction process, find out the weighted distance between corrupted and corrected exemplar as

$$\Phi = \left[\sum_{i=1}^{n} \left| \delta_i' - \kappa_i' \right| \right] \tag{13}$$

If $\Phi = 0$ then distance becomes zero and engine decides that the equilibrium point is reached, i.e., corrupted bits in decryption process are safely recalled by hetero-correlators.

If $\Phi \le \rho$, then assign corrected bits to $\delta, i.e., (\delta = \kappa), (\delta' = \kappa')$ and perform step 2 until distance of exemplar becomes zero.

If $\Phi > \rho$, then the engine confirms that adversary does the correction process, therefore system has been terminated.

The ρ is a vigilance parameter and it is calculated as $\rho = (n - \text{mod}(n,2))$ i.e., $0 \le \rho \le (n - \text{mod}(n,2))$ and n represents number of bits in an exemplar. The parameter ρ provides minimum energy for the bits correction between genuine subject and partially corrupted cipher bits and also it prevents local minima of the system. This parameter also gives more complexity to the impostor to get back the original messages.

Finally, recalled bipolar bits are converted to its equivalent binary bits. These sequences of corrected bits represent the original bits. The number of error bit recovery is based on ρ and E parameters. If 7-bit exemplar is used, then the parameters $\rho = 6$ and $E = \{2,3,2\}$ provide a better result in the error correction process.

V. Experimental Results

The proposed approach has been implemented and results were analysed. Efficacies of SIC and NRIC have been evaluated. The NRIC system's time complexity was measured, in that there were no recalling processes involved since the encrypted bits were decrypted by the enrolled iris key. Hence its enciphering and deciphering process depends on the time complexity of iris-matching algorithm.

Next, the performance of the NRIC system was measured by computing the time complexity of auto and hetero-correlators' recalling and encryption/decryption processes. In the next experiment iris key energy complexities was calculated in the case of cracking the messages by the impostors. Finally, the guarantee issues of getting back original bits were evaluated with respect to the energy variation of auto and hetero-correlators. The detailed description of each experiment is discussed in the following sections.

A. Speed performance

Time complexities of encryption and decryption process have been evaluated for the SIC system. In that decryption process required more time than encryption process, since the decryption was performed after extracting and matching the iris features at one time. The complexity of iris matching algorithm was dependent on the size of the iris keys present in the system.

The complexity of searching iris keys iris key matching system with linear search is O(N) and with binary search is O(log N). The NRIC system required slightly more time than the SIC approach because of its error correction engines require more time to predict the original patterns from the partially corrupted patterns. The search time of encryption and decryption processes of SIC and NRIC are illustrated in Fig. 4.

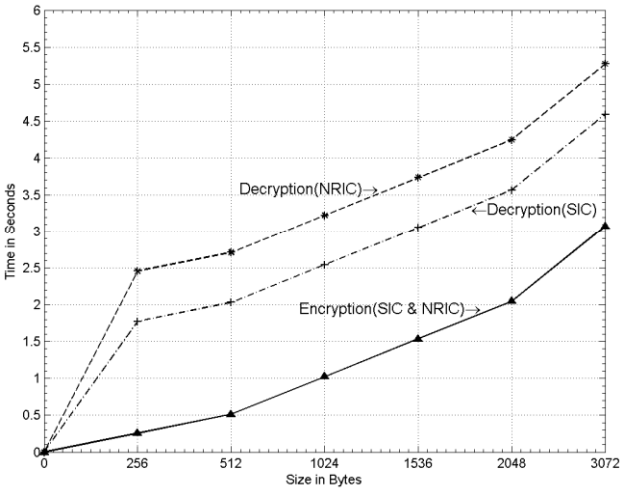


Figure 4. Encryption and decryption time complexity of SIC and NRIC.

B. Recalling time

The recalling time of auto and hetero correlations were dependent on size of the connection matrix in the error correction process. The connection matrix was formed based on the number of bits processed by the cipher text. In accordance with the number of patterns and bits per exemplar, the recalling time of auto and hetero correlators were evaluated and shown in Fig. 5.

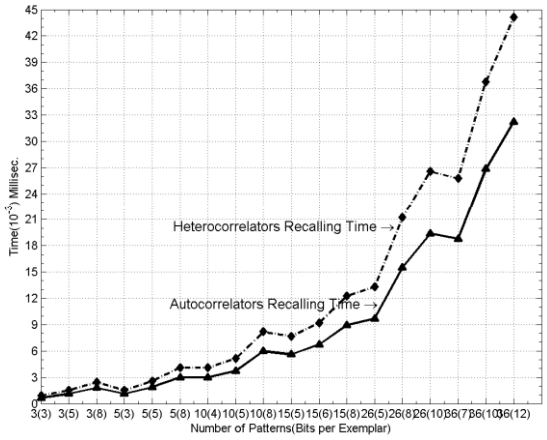


Figure 5. Auto and hetero correlators' recalling time.

C. Performance issues

The guarantee issue of recalling process for correlators was associated with two factors such as connection matrix of the error correction engine and artifacts occurring on the iris patterns. It provides nearly 97% of recalling entire pair of trained patterns because of its local minimum of the energy surface. However, in this paper, vigilance parameter was used to put off local minimum attained by the system, i.e., energy for the bits correction in between genuine subject and partially corrupted cipher bits were computed to prevent the local minima of the system. This parameter also produced more complexity to the impostor to get back the original messages. The factors of artifacts are fully concerned with three possessions such as acquisition time users' co-operation, non-iris fractions occurring on iris and artifacts emerging in the core area of iris. The guarantee issues of error correction process for auto and hetero correlators are based on number of patterns and bits per patterns used in the error correction process.

The guarantee performance of recalling process was evaluated based on the Hamming distance between the corrected bits and trained pairs. Multiple training was used to recall several patterns. In this training, if pattern was not recalled by the connection matrix by satisfying vigilance parameter then train the patterns again by changing energy constants, form a new connection matrix and performing recalling process. This process was repeated until recalling entire patterns by checking vigilance parameter. However, trained patterns require sufficient number of bits to increase the percentage of accuracy. Figure 6 shows the accuracy of recalling patterns using auto and hetero correlators.

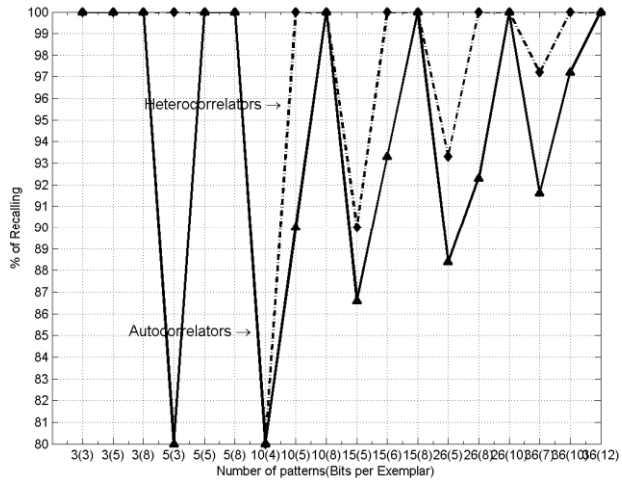


Figure 6 Accuracy of recalling patterns using auto and hetero correlators.

D. Impostor complexity

The probability of the presence of errors in the non-repudiation process was assessed based on the number of bits variation. These variations occur due to the environment, illumination, occlusion of eyelids/eyelashes and other artifacts. In this experiment, the number of bits corrupted in different sessions was studied and verified in which situations brute force search by an intruder can crack the iris crypto key. For the experiment, different eye images were captured at different sessions from the same subjects and their changes measured. Figure 7 illustrates the error bit variation in different criterion. The changes in bits may not be stable for all kind of capturing because due to diverse changes the random alteration of bits was assorted. The efficiency of the iris cryptosystem was evaluated in accordance with key stability and strength. The strength of the key can be evaluated based on entropy principles. If message source alphabet was $A = \{a1, a2\}$ and the symbol probability P (a_1) = 0.088 and P (a_2) = 0.103 then the entropy of the source symbol was 0.6495 bits/symbol. If an intruder can tap the message, the probability of retrieving the original message was ranged from 0.2 to 1 based on the error

bits of iris code. That is, if n bits were error then 2^{n-26} times of complication for brute force search was made to an intruder. Thus the retrieving of the original messages has been made complicated to the impostors. It provided a high key strength for any cryptography system. This key cannot be stolen or missed and gave more stability to the cryptosystem. These types of bio keys can be produced every time the users want to communicate secretly at non-secure channels. In addition, experimental results show that this approach could easily be adopted in the on-line cryptography systems as well.

E. Re-enrolments

Another design issue of integrating biometrics with cryptography is the re-enrolments because biometric cryptosystem is a reliable alternative for password protection while releasing or direct usage of biometric key as a cryptography key.

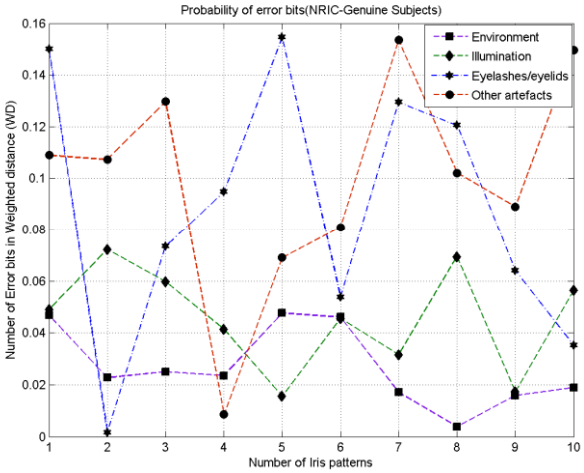


Figure 7. Error bit variation for the same subject in different criterion.

Hence encryption algorithm needs efficient solutions, which are periodically updated biometric templates. Thus user can register their patterns once in a month or other period of time maintained in the system. Since some of the system exploits biometric key for safeguarding mathematical cryptographic keys or others may utilize as a part of the biometric template. Nevertheless if biometric databases are permanently stored in the local workstation for a period of time, which is not secure, a system should employ the recently enrolled iris keys for encryption process that increases the system security and avoids eavesdropper attacks than the lifelong biometric templates.

Thus the iris-based cryptosystem performs better accuracy by using re-enrolments. In this paper, subjects' iris patterns were periodically enrolled once in a week in order to measure the stability of the iris keys. However the keys variation weighted distance was ranging from 0.0 to 0.19. This range was fixed by statistical measures of iris recognition algorithm. Thus these random variations were due to artifacts or other non-iris sources. However the periodic amendment of genuine subjects' iris key produced more brute force search to the impostors than the ordinary system.

VI. CONCLUSION

This research paper suggests a novel approach for iris based cryptography system. The crypto keys have been generated using iris patterns, which is stable throughout a person's lifetime as well. Its inter-class variability for a person is very large since it creates more complexity to crack or guess the crypto keys. This approach has reduced a complicated sequence required to generate keys as in the traditional cryptography system. It can also generate more complex iris keys with minimum amount of time complexity, which is aptly suited for any real time cryptography system. This resolves the key repudiation problem occurring in the traditional system. The hetero-correlators can predict the number of bits corrupted in the decryption process with the help of vigilance parameter. The performance of the proposed approach is found to be satisfactory.

In near-future, multi-modal cryptosystem will be suggested to integrate biometric template to increase degree-of-security in the non-secure data transmission.

REFERENCES

[1] Albert Bodo, 'Method For Producing a Digital Signature with Aid of a Biometric Features', German patent DE 42 43 908 A1, 1994.

[2] Davida G.I., Frankel Y. and Matt B.J., 'On enabling secure applications through off-line biometric identification', Proc. of IEEE Symposium Privacy and Security, Oakland, California, USA, pp. 148-157, 1998.

[3] Davida G.I., Frankel Y., Matt B.J. and Peralta R., 'On the relation of error correction and cryptography to an offline biometric based identification scheme', Proc. Workshop Coding and Cryptography (WCC'99), PARIS (France), pp. 129-138, 1999.

[4] Linnartz M.G. and Tuyls P., 'New Shielding Functions to Enhance Privacy and Prevent Misuse of Biometric Templates', AVBPA 2003, Guildford, UK, pp. 393-402, 2003.

[5] Clancy T., Kiyavash N. and Lin D.J., 'Secure Smartcard-Based Fingerprint Authentication', Proc. of ACM SIGMM workshop on Multimedia, Biometric Methods and Applications, New York, USA, pp. 45-52, 2003.

[6] Monrose F., Reiter M., Li Q. and Wetzel S., 'Cryptographic key generation from voice', Proceedings IEEE Symposium on Security and Privacy, Oakland, California, pp. 201-213, 2001.

[7] Yao-Jen Chang, Wende Zhang and Tsuhan Chen, 'Biometrics-Based Cryptographic Key Generation', IEEE International Conference on Multimedia and Expo, Taipei, Taiwan, (0-7803-8603-5/04), pp. 2203-2206, 2004.

[8] Uludag U., Sharath Pankanti, Salil Prabhakar, Anil K. Jain, 'Biometric Cryptosystems: Issues and Challenges', Proceedings of the IEEE, Vol. 92, No. 6, pp. 948-960, 2004.

[9] Daneil Schonberg and Darko Kirovski, 'Iris compression for Cryptographically Secure Person Identification', Proceedings of IEEE Data Compression Conference (DCC'2004), Snowbird, UT, USA, pp. 459-468, 2004.

[10] Feng Hao, Ross Anderson and John Daugman, 'Combining cryptography with biometrics effectively', Technical report of University of Cambridge, No. 640, pp. 3-17, 2005.

[11] Bremananth R and Chitra A, 'An efficient biometric cryptosystem using autocorrelators' International Journal of Signal Processing 2;3, pp.158-164, 2006.

[12] R.Bremananth and A.Chitra, "A novel approach for high authentication based on Iris keys", World Scientific and Engineering Academic Society Transaction on Information science and applications, Issue 9, Vol. 2, pp.1420-1429, 2005.

[13] R.Bremananth, A.Chitra, "A new methodology for person identification system", Sadhana, Indian academy of Sciences, Vol.31, Part 3, pp.259-276, 2006.

[14] R.Bremananth, A.Chitra, "Rotation Invariant Recognition of Iris", Journal of Systems Science and Engineering, Vol.17, No.1, pp.69-78, 2008.

[15] Bart Kosko, 'Bidirectional Associative Memories', IEEE Transactions on systems, Man, and Cybernetics, Vol. 18, No. 1, pp. 49-60, 1988.

[16] Yeou-Fag Wang, Jose B. Cruz and James H. Mulligan, 'Two coding strategies for bidirectional associative memory', IEEE Transactions on Neural networks, Vol. 1, No. 1, pp. 81-92,1990.

[17] Yeou-Fag Wang, Jose B. Cruz and James H. Mulligan, 'Guaranteed recall of all training pair for bi-directional associative memory', IEEE Transactions on Neural Networks, Vol. 2, No. 6, pp. 559-567,1991.

[18] P.Arul, A.Shanmugam, Generate a Key for AES using Biometric for VOIP Network Security, Journal of Theoretical and appllied Information Technology, Vol.5, No. 2, pp. 107-112, 2009.
(http://www.jatit.org/volumes/research-papers/Vol5No2/2Vol5No2.pdf)

AUTHORS PROFILE

Bremananth R received the B.Sc and M.Sc. degrees in Computer Science from Madurai Kamaraj and Bharathidsan University in 1991 and 1993, respectively. He obtained M.Phil. degree in Computer Science and Engineering from GCT, Bharathiar University, in 2002. He received his Ph.D. degree in 2008 from Department of Computer Science and Engineering, PSG College of Technology, Anna University, Chennai, India. He has completed his Post-doctoral Research Fellowship (PDF) from the School of Electrical and Electronic Engineering, Information Engineering (Div.) at Nanyang Technological University, Singapore, in 2011. He has 18+ years of experience in teaching, research and software development. Currently, He is an Assistant Professor in the Information Technology department, Sur University College, Sur, Oman, affiliated to Bond University Australia. He received the M N Saha Memorial award for the best application oriented paper in 2006 by Institute of Electronics and Telecommunication Engineers (IETE). His fields of research are acoustic holography, pattern recognition, computer vision, image processing, biometrics, multimedia and soft computing. Dr. Bremananth is a member of Indian society of technical education (ISTE), advanced computing society (ACS), International Association of Computer Science and Information Technology (IACIT) and IETE. He can be reached at bremresearch@gmail.com.

Ahmad Sharieh had two bachelor degrees: one in Mathematics and one in Computer Sciences. He had master degree in Computer Science and High Diploma in Teaching in Higher Education. He had PhD in Computer and Information Sciences from Florida State University 1991. Sharieh worked as Assistant Professor in Fort Valley College / USA and The University of Jordan (UJ) / Jordan. He worked as Associate Professor with Amman Arab University for Graduate Studies (AAUGS) / Jordan and The University of Jordan. He worked as Dean of King Adbullah School for Information Techology/Jordan. Currently, he is a professor of Computer Sciences and Dean at Sur University College (SUC), Oman. He published articles in journals (27), in conferences (22), and authored and prepared books (14). He gained grant for eight research projects from UJ and Europe. He developed several software systems such as: Teaching Sign Language, e-learning Modeling and Simulation, and Online (Automated) Exams. He is on the editorial board of several journals and conferences and a referee of several others. His research areas are Distributing Systems, Expert Systems, E-Government, E-Learning, Parallel Processing, Pattern Recognition, Software Engineering, Wire/Wireless Communication, Modeling and Simulation.

On the transmission capacity of quantum networks

Sandra König
Alpen-Adria Universität Klagenfurt
9020 Klagenfurt, Austria

Stefan Rass
System Security Research Group
Alpen-Adria Universität Klagenfurt
9020 Klagenfurt, Austria

***Abstract*—We provide various results about the transmission capacity of quantum networks. Our primary focus is on algorithmic methods to efficiently compute upper-bounds to the traffic that the network can handle at most, and to compute lower-bounds on the likelihood that a customer has to wait for service due to network congestion. This establishes analogous assertions as derived from Erlang B or Erlang C models for standard telecommunications. Our proposed methods, while specifically designed for quantum networks, do neither hinge on a particular quantum key distribution technology nor on any particular routing scheme. We demonstrate the feasibility of our approach using a worked example. Moreover, we explicitly consider two different architectures for quantum key management, one of which employs individual key-buffers for each relay connection, the other using a shared key-buffer for every transmission. We devise specific methods for analyzing the network performance depending on the chosen key-buffer architecture, and our experiments led to quite different results for the two variants.**

Keywords-Quantum network; Quantum cryptography; network transmission capacity; queuing network; system security.

I. INTRODUCTION

It took about two decades ever since quantum key distribution (QKD) has been proposed by [1] (the famous BB84 protocol) until the first experimental implementations of a quantum network were presented by the DARPA [2] and the European Union [3]. While the theory behind secure key-delivery between Alice and Bob is well-understood (see e.g. [4] for a proof regarding the security of BB84), the theory of network design and performance analysis has apparently seen rather limited attention over the last years. The works of [5] and [6] both considered the design of a network from the topological point of view, and in terms of optimal security and performance. In this work, we go the other way, asking for the best performance that we can get from a given quantum network infrastructure. In particular, we provide algorithmic means to answer two questions:

1. What is the maximal transmission capacity achievable in the network (using any classical routing scheme)?
2. What is the likelihood of local congestion that would temporarily disconnect the (logical) channel between any two peers in the network?

The second question can be rephrased into asking how likely a customer is to wait when asking the network for a secure delivery of payload from one point to another.

Our results are hence related to the field of communication theory, channel capacity and network coding. Particularly the latter has led to valuable insights (cf. e.g. [7], [8], [9]) regarding the rate at which information can be send through the network. Contrary to these (and many other related) approaches, we do not employ classical information theory to quantify the capacity, but rather work with the directly known performances of each link in the network. Similar to our work, network outage probabilities are as well discussed in [10], [11], [12] and [13], where most research effort, as it seems, has been put on wireless networks. So far, the problem appears hardly considered in (hard-wired) networks or quantum networks. In the quantum computing domain, the work of [14], [15], [16] and [17] is closely related. Contrary to ours, however, it strongly relies on quantum techniques and is less focused on algorithmic methods to analyze a given infrastructure. The work of [16] is particularly interesting as it employs percolation theory (which is rarely used in the related literature). The problem is studied elsewhere in [18], which comes up with proposals on how to enhance the existing capabilities once they are known. Here, we work out the limits similarly to the Erlang B and Erlang C models, so as to be able to improve them based on this related research. In the wireless domain, the interesting work of [19] deals with spread-spectrum techniques and uses Poissonian processes for determination of the network capacity, but is specific for this particular encoding technique. We explicitly avoid such restrictions here, but adopt some assumptions on the quantum key-management models (following the proposals of [20]; cf. also [21] for another discussion related to quantum key- and network-management). Finally, we mention the work of [22], who attempts to solve the problem of end-to-end quantum communication using a three-party protocol. This architecture is essentially different from what has been implemented nowadays, and thus subject of future considerations.

Among the sales arguments for quantum key distribution is its capability of running over existing fibre-optic lines. This claim has been substantiated in the demonstration of the SECOQC- and DARPA-Networks [2], [3]. Hence it is fair to assume that the topology of the network is fixed and that the individual key-generation rates on each link are known (cf. [3] for examples). Moreover, following the so-far proposed architectures of relay devices in a quantum network (see [2], [3],and [20]), it is reasonable to consider a quantum network as a system of interconnected buffers, where the secret message is repeatedly decrypted and re-encrypted before forwarding it to the next hop. The key-buffers in each relay node are constantly refilled by QKD protocols running in the background 24/7 and

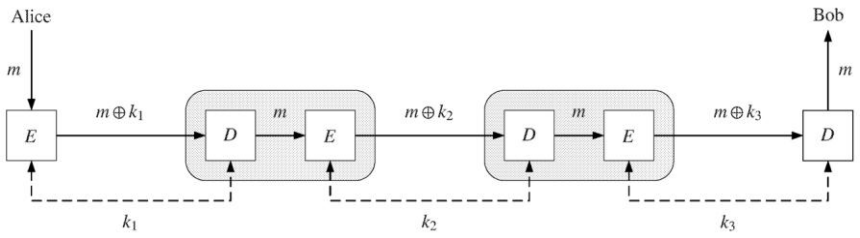


Figure 1 Trusted relay with re-encryption in each hop

endlessly generating key-material for later usage. This transmission regime, in its simplest form, is known as trusted relay, and is widely used in nowadays demonstration networks (cf. [2] and [3]). A transmission of a message m from Alice to Bob along a sequence of nodes that share keys $k_1, \dots, k_3$ is displayed in Figure 1.

Organization of this work: in Section II, we describe the graph- and queuing model (Section II.A) used to analyze a quantum network. In particular, we will use maximal flows (briefly introduced in Section II.B) to compute bounds on the payload that the network can handle. Section III is divided into two main parts, giving algorithms for computing end-to-end transmission capacity (Section III.A) and waiting probability if a chosen path through the network is blocked due to congestion (Section III.B). In both cases, we show how to use standard maximum flow and shortest path algorithms to compute the desired quantities easily and efficiently from the known link capacities in the underlying graph model. The process is illustrated by an example (Section III.C), before conclusions follow in Section IV.

II. Mathematical Model

We will not burden ourselves with the details of any particular quantum key distribution (QKD) facility, but restrict our attention to the following model of a quantum network (QNet): let a QNet be modeled as an undirected graph $G = (V, E)$ with adjacent nodes sharing secret keys thanks to QKD. That is, on any line $u - v$ (with $u, v \in V$) maintain key-buffers on either side to store QKD key material for subsequent transmissions. These key-buffers are nothing else than queues, in which key-bits are inserted on a deterministic basis (we assume the QKD-devices to generate key-bits at constant rate). key-bits are used up on a random basis, depending on incoming payload for secret transmission. Based on this view, we can cast the QNet into a standard queuing network.

A. Quantum Networks as Queuing Networks

An open queuing network is a system in which a customer enters the network at some node, and moves onwards through the links, where he occasionally has to wait (queue) until he is served at the next node (i.e. he can enter the next node). Central questions in queuing theory regard the average time to wait until the customer reaches his destination point, or the average number of customers lining up in any given queue (link) in the network. For a quantum network, we can equally well set up such a model, based on the following correspondence:

1. incoming customers equal newly generated key-bits
2. leaving customers equal the (one-time) use of key-bits for Vernam-encryption of messages
3. a queue equals a key-buffer (storing bits for subsequent usage, or equivalently, hosting customers for subsequent service)

Observe that the generation of key-bits is deterministic, while the arrival of messages is non-deterministic. If we consider the QNet as a backbone network, then it is reasonable to consider the event of an incoming message bit as a Poisson-distributed random variable. In Kendall-notation, the link $u - v$ in a QNet therefore is nothing else than a $D/M/1$-queue, disregarding physical size limits of the key-buffers for now. The graph G modeling the QNet thus constitutes a network of queues, and we are interested in its stationary distribution (so it exists).

Remark: an alternative view yielding equivalent results is by associating incoming payload bits with customers, who get served (encrypted) based on how much key-material is available in the buffer. In this case, we would have to think of the message-queue as the physical incarnation of the queue model under consideration. For simplicity, and because real-life implementations work with a key-buffer too, we shall consider the first of these two views, keeping the second view in mind whenever needed.

Sufficient conditions for the existence of stationary distributions in queuing networks are well-known, such as Jackson's theorem for Jackson-networks or one of its generalizations, such as the BCMP-theorem. Openness of the network is assured, since the graph G is a mere transportation medium and messages necessarily leaving the network at some stage. Still, we cannot make any generally valid assertions on the routing algorithms implemented within the system. In lack of such information, we will try to find upper bounds to the transmission capacity by invoking maximum flow theory. This has the additional advantage of our results applying to conventional routing as well as network coding approaches for transmission.

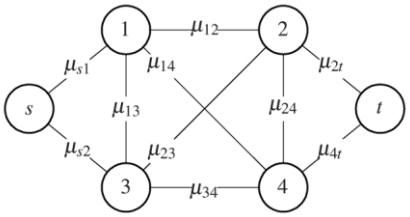


Figure 2 Network with link performances

B. Flows

To illustrate the approach, consider the example network topology shown in Figure 2. Assume that after start-up, all key-buffers are empty and the QKD-protocol on link $v_i - v_j$ start producing key-material at constant rate μ_{ij} per time unit. It follows that after one unit of time, the maximal transmission capacity of the network is determined by the maximum $s-t$-flow, constrained by the existing key-material on each link. As all links regenerate key-bits at constant rate, the minimum cut will not change over time. Let $C \subseteq E$ be the minimum edge cut, then the maximum flow has capacity $c(f_1) = \sum_{(v_i,v_j)\in C} \mu_{ij}$ after the first unit of time. After t units, we have the capacity $c(f_t) = \sum_{(v_i,v_j)\in C} t \cdot \mu_{ij} = t \cdot c(f_1)$. The consumption of key-bits happens upon arrival of payload to be transmitted secretly from the sender s to the receiver t. Hence, we can consider the entire network as one large queue, whose internal servicing is done by routing, network coding, or otherwise. From outside, we have $c(f_1)$ as the deterministic rate at which key-bits arrive for later consumption, and we are back at the $D/M/1$-queue. Considering multiple access-points to the network is trivial by switching to a multi-source-multi-sink flow. Unlike standard queueing disciplines, optimality in our context means the incoming amount of key-material outweighing the arrival of messages, i.e. an "unstable" queue whose expected length is infinite

C. Key-Buffer Architectures

It is easy to set up the devices so as to realize a single-path transmission as illustrated in Figure 1. However, it would not be reasonable to assume nodes to have only two ports, so the internal management of quantum keys is slightly more involved. Going back to Figure 1, we can instantly fix the problem of the message popping up in plaintext within each relay node by simply XOR-combining both, the incoming- and outgoing key into a single "relay-key" [20]. Figure 4 illustrates the idea: for the relay from node A to node B over node R, the latter would XOR-combine ($\oplus$-operation) k_A and k_B into k_{AB}. Consequently, we would only store k_{AB} in an *individual buffer* for this link (see Figure 3; right). If the relay is trusted, then it may alternatively decrypt the incoming message and re-encrypt it before passing it onwards (as shown in Figure 1). Consequently, we would have to maintain *shared buffers* for each I/O-port, as displayed in Figure 3 (left) .

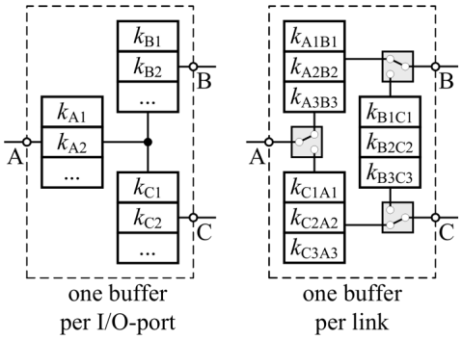


Figure 3 Shared vs. individual key-buffers

III. Results

We are now ready to present our main results.

A. End-to-End transmission capacity

Given the (constant) rate μ of key-bits generated on link $u - v$, we can ask for the maximal average rate λ of arriving messages that we are able to encrypt. Or stated differently: if we know the service rate, what is the highest rate of incoming customers that we can handle? Obviously, a necessary condition is the rate of arriving customers not exceeding the service rate. This is almost sufficient, as the following result shows:

Proposition 1. Let a $M/D/1$-queue be given and denote the average arriving rate by λ. For a given constant service rate μ, any arrival rate $\lambda < \mu$ leads to a stable system.

So although the case $\lambda = \mu$ may be fine for a system with deterministic arrival, it is not appropriate for random systems.

Proof: Denote by N the number of customers in the system and by $p_n = \Pr\{N = n\}$ the probability of being in state n. In a stable situation, the total rate out of this state n equals the sum of the incoming rates from states $n-1$ and $n+1$, which yields for every $n \in \mathbb{N}$,

$$\lambda\, p_{n-1} + \mu\, p_{n+1} = (\lambda + \mu) p_n.$$

This recursive formula can be rewritten as

$$p_{n+1} = \left(\frac{\lambda}{\mu}\right)^n \cdot p_0,$$

and since probabilities sum up to one we get

$$1 = \sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n \cdot p_0.$$

This is simply a geometric series, which converges if and only if $\lambda < \mu$, as stated. □

The above proof also shows that the probabilities p_n can be calculated via $p_n = \rho^n(1-\rho)$ for any $n \in \mathbb{N}$ (by using $1 = \frac{p_0}{1-\rho}$), where $\rho = \frac{\lambda}{\mu}$.

Putting this to practice within a quantum network is straightforward in two steps:

1. Upon given key-generation rates on each link, use

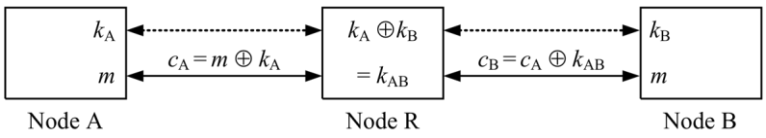


Figure 4 Forwarding with link-specific keys

these rates as edge-weights in an undirected graph and determine a maximal flow from the source node to the sink node. Call the value of this flow μ.

2. By Proposition 1, a payload of up to $\lambda < \mu$ bits per time unit can be handled by the quantum network in a perfectly secure manner (assuming trusted relay).

B. Waiting probability

Here, we need to distinguish two architectures in our theoretical considerations: let a node $v \in V$ be given, whose neighbors are $u_1, u_2, \ldots, u_{\deg(v)}$, where $\deg(v)$ denotes v's degree. Either each route passing through v is associated with its individual key-buffer (perhaps via logically partitioning the overall key-material somehow; cf. the right side of Figure 3), or all incoming and outgoing flow draws from the same key-buffer (Figure 3; left), in which case a very busy line can affect the capacities of other routes through v. However, short-term traffic peaks are easier to handle with this architecture. We consider both variants separately.

Individual key-buffers

To estimate the probability that one has to wait to get the key-bits needed for encryption anywhere along its way from the sender to the receiver, we first focus on the waiting probability in one particular node w. Observe that since the key-buffers are not shared, distinct links from a node w to any of its neighbors act independently. So we can restrict our considerations to any (arbitrary) key-buffer within w. Notice, however, that a link from w to its neighbor v has to be treated differently than the link from v to w, since we are concerned with *forwarding* packets.

Let $v \in V(G)$ be any node, and denote its neighbors by $N(v)$. Pick any key-buffer within v that refers to the connection $v - w$, where $w \in N(v)$. Let the incoming traffic per time unit on the route from v to w be $Poisson(\lambda)$ distributed, and assume the QKD protocol between v and w to reproduce an amount of ν bits per time unit. Finally, assume the key-buffer to be full at the beginning. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables $X_i \sim Poisson(\lambda)$ where X_i is the traffic at time unit i. The corresponding filling level of the key-buffer at time i can never exceed the capacity L and is thus given by

$$Y_i = \min\left\{L, L + (i-1)\cdot\mu - \sum_{j=1}^{i} X_j\right\}, \qquad (1)$$

assuming that we start off with the full key-buffer and re-fill it at rate μ after the first time-unit (i.e. we do no refill within the first time-unit because the buffer is full already).

We are interested in the probability for the link being blocked, i.e. the likelihood of an empty key-buffer at time unit t. Hence, we ask for $p(t) := \Pr\{Y_t = 0\} = \Pr\{Y_t \leq 0\}$. From $L > 0$ we deduce $\min\{L + (i-1)\cdot\mu - \sum_{j=1}^{i} X_j, L\} \leq 0$ if and only if $L + (i-1)\cdot\mu - \sum_{j=1}^{i} X_j \leq 0$. Hence,

$$p(t) = \Pr\left\{L + (t-1)\mu - \sum_{j=1}^{t} X_j \leq 0\right\}$$

$$= \Pr\left\{\sum_{j=1}^{t} X_j \geq L + (t-1)\mu\right\}$$

If the traffic load over different time-units is independent, then $\sum_{j=1}^{t} X_j \sim Poisson(t\cdot\lambda)$ so that the above probability boils down to a mere evaluation of the Poisson distribution function $F(x|\lambda) = \sum_{i=0}^{\lceil x\rceil-1} \frac{\lambda^i}{i!} e^{-\lambda}$ and comes to

$$p(t) = 1 - F(L + (t-1)\mu | t\cdot\lambda).$$

It is legitimate to ask what happens if the refilling of the key-buffer happens on a random basis as well. Call X'_i the amount of fresh key-material in time-unit i. We can easily replace the term $(i-1)\cdot\mu$ in (1) by $\sum_{i=1}^{t-1} X'_i$ so as to take this randomness into account, but the distribution of $X'_{i+1} - X'_i$ is no longer Poissonian (mostly because the difference is not bounded from below). A straightforward way out of this dilemma is considering Gaussian approximations to the Poissonian densities, which takes us back to the wonderful world of distributions closed under convolution. In other words, if we approximate $X_i \sim Poisson(\lambda)$ by $\tilde{X}_i \sim \mathcal{N}(\lambda, \lambda^2)$, the above derivation and result becomes obvious. We leave this track aside here and go back to the deterministic refilling, giving us the following result:

Proposition 2. Let $w \in V(G)$ have neighbors $N(w)$, and consider the key-buffer shared with an arbitrary but fixed neighbor $v \in N(w)$ of w. Denote by L_v the size of the key-buffer associated with the link $w - v$ and assume that new key-bits are generated at a constant rate μ. The number of incoming message bits from w to v is assumed to be $Poisson(\lambda)$ -distributed. If $\lambda > \mu$, then the probability of waiting within w during a transmission is

$$\begin{aligned} p_{w\to v} &= \Pr\{\text{forwarding from } w \text{ to } v \text{ gets delayed}\} \\ &= 1 - F(L_v + (t_0 - 1)\mu | t_0\lambda) \\ &= 1 - \sum_{i=0}^{[L_v + (t_0-1)\mu]-1} \frac{(t_0\lambda)^i}{i!} e^{-t_0\lambda}, \end{aligned}$$

where $t_0 = \frac{L_v}{\lambda - \mu}$.

Proof: The argument is merely a matter of noticing that it takes a period of $t_0 = \frac{L_v}{\lambda-\mu}$ time units to entirely empty the key-buffer, if λ bits are used up with μ bits growing back per time unit (the time for encryption is considered negligible). Hence, the average expenditure is $\lambda - \mu$. □

The alert reader might utter concerns about the stochastic independence of incoming traffic over different time-units. There are (at least) two ways to justify this assumption:

- Considering the transmission as a process resting on various routing protocols, we can reasonably assume the network's routing regime to rearrange, encode and decode, and to partition the messages in a way that stochastic correlations between packets are negligible. Notice that this in no way rules out the possibility of linking packets to each other via sequence numbers or matching delivery addresses. However, this "weaker" type of correlation does not necessarily imply dependencies among the payloads of different packets, e.g. when a long sequence of independent cryptographic key-material is transmitted.
- In case that the network is merely used for continuous shared key establishment (in fact, this is the way in which a quantum link is generally supposed to be used [3]), we can safely assume incoming traffic packets as stochastically independent, for otherwise we would have interdependence among key-bits. This is undesired for cryptographic keys, particularly for quantum keys (as it reduces the key's entropy).
- While Proposition 2 refers to only a single node, it is more interesting to find out how likely a blockage along a path from any node to any other node is. In the model of individual key-buffers, this problem boils down to identifying a path whose blockage probability is minimal. We can simply invoke any shortest-path algorithm for that matter, if we assume blockages to happen *independently* of each other. Consider a node $w \in V(G)$ having neighbors $N(v)$.

Observe that Proposition 2 is concerned with the likelihood of blockage when *forwarding* a message from w onwards. Hence, we need to cast the undirected network model graph into a directed graph by converting an undirected edge into two directed edges (with opposite directions).

Each link $w \to i$ for $i \in N(v)$ maintains its own key-buffer with blocking probability $p_{w \to i}$ as given by Proposition 2. We transform the undirected graph $G = (V, E)$ into a weighted directed graph $G' = (V', E', c)$ such that

1. Each link $\{u, v\} \in E(G)$ is carried over into two links $(u, v), (v, u) \in E'$ with cost $c(u, v) = c(v, u) = 0$.
2. Each node $v \in V(G)$ having a number $n = |N(v)|$ of neighbors is expanded into a complete graph with n nodes $v_1, \dots, v_n$, each of is connected by two directed edges in either direction. Each edge (v_i, v_j) is added to E' with the cost $c(v_i, v_j) = -\log(1 - p_{v_i \to v_j})$ according to Proposition 2. The set of edges joining v to its neighbors in G, i.e. the set $\{\{v, u\}: u \in N(v)\} = \{(v, u_1), (v, u_2), \dots, (v, u_n)\}$ is carried over to E' as $\{(v_1, u_1), (u_1, v_1), \dots, (v_n, u_n), (u_n, v_n)\} \subseteq E'$ with weights all zero.

Figure 5 illustrates this transformation.

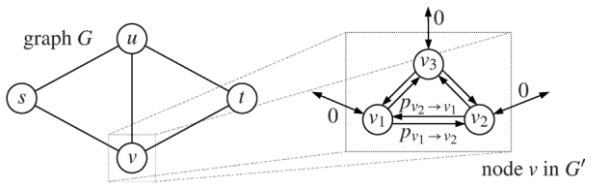


Figure 5 Transformation of a node with individual key-buffers

On G', we can run any shortest-path algorithm, to determine the minimum likelihood of blockage using single-path routing. For any given sender s and receiver t, let their most reliable interconnecting path have "weight" p in G'. Then, regardless of the routing, we have

$$\Pr\{\text{message will be blocked}\} \geq 1 - \exp(-p),$$

because p is the weight of the *shortest* path in G', i.e. the *most reliable* path in G. No matter what the routing actually does, it cannot do better than choosing the best path possible, hence the value

$$1 - \exp(-p) = \Pr\{\text{at least one node is blocked}\}$$

is a lower-bound to the actual likelihood.

Shared key-buffers

In the case of shared key-buffers we assume the incoming flows from different nodes to be independent. Then the distribution of the total flow trough $w \in V(G)$ at time t follows a Poisson distribution with parameter $\lambda = \sum_{i \in N(w)} \lambda_i$, where λ_i denotes the incoming traffic flow from node w to its neighbor i. Similarly, all neighboring links $i \in N(w)$ contribute μ_i bits of fresh key-material to the common key-buffer, giving a total refreshing rate of $\mu = \sum_{i \in N(w)} \mu_i$. With λ, μ we can invoke proposition Proposition 2 again to calculate the probability of a node being blocked in this case.

A little more care is to be taken when asking for the chance of blocking somewhere across the network as a whole. In this case, we use a transformation that is normally used to calculate maximal flows with vertex capacities. The transformation from the undirected graph $G = (V, E)$ (see Figure 6a for an example) to the directed weighted graph $G' = (V', E')$ is now specific for a sender s and receiver t, and proceeds as follows (cf. [23]):

1. Each node v including s and t is replaced by two nodes $v_{\text{in}}, v_{\text{out}} \in V'$, and a directed edge from v_{in} to v_{out} is placed to E'. This link $v_{\text{in}} \to v_{\text{out}}$ gets assigned the cost $-\log p$, where p is the blocking probability calculated as described above.
2. Each undirected edge $(u, v) \in E$ is replaced by two directed edges $u_{\text{out}} \to v_{\text{in}}$ and $v_{\text{out}} \to u_{\text{in}}$. See Figure 6b and Figure 6c for an illustration.
3. The nodes s_{in} and t_{out} are deleted, as well as all edges going into s_{in} and out of t_{out}.
4. Those nodes who remain to be assigned a cost receive zero cost. The resulting graph is shown in Figure 6d.

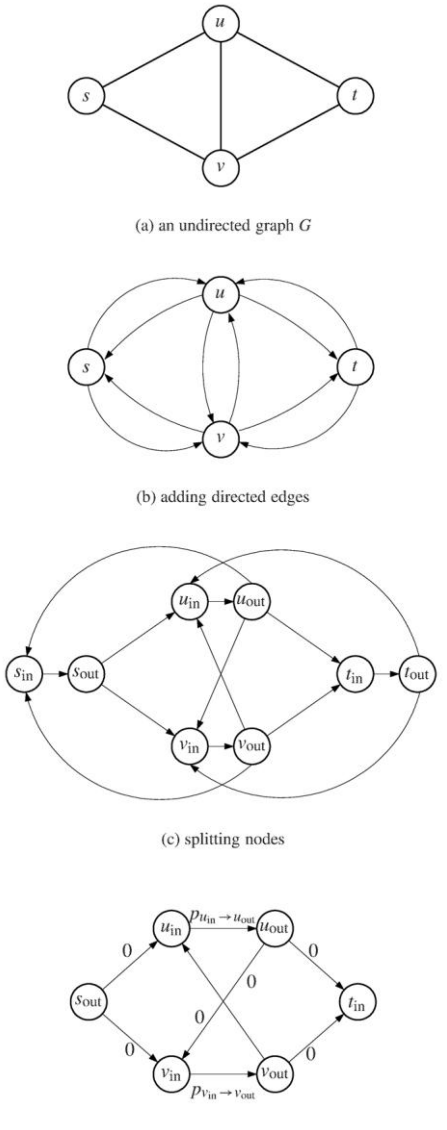


Figure 6 Transformation for shared key-buffer

Once having found the shortest path in G' between s and t, we can draw exactly the same conclusion as above: if p is the weight of this path in G', then the chance of this path being blocked for *any* path-based routing-scheme is lower-bounded by $1 - e^{-p}$.

C. Example

To get a more intuitive understanding of the above results, we give a simple example. Consider a modified version of the graph from Figure 2, with six nodes but with neither an edge between vertices 2 and 3 nor between 2 and 4. We call node 0 the sender and the receiver shall be node 5 (cf. Figure 7). We

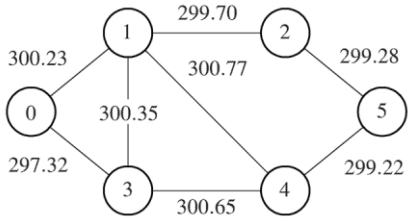


Figure 7 Example network (link performances $\boldsymbol{\mu}$ shown in kbit/sec)

let L be 5000 kbits and choose the rates μ (at which new key-material is produced) randomly between 280 and 320 kbits and let the rate of the incoming message bits be $\lambda = \mu + a$ (see Table 1(a)), where a is a positive constant ($a = 5$ in this specific example). Under this setting, we get the average probabilities $p_{v\to w}$ shown in Table 1(a) for the incident of waiting between v and w, where the average was taken over $N = 100$ calculations.

Using the transformation described above we get that the probability of getting stuck is lower bounded by 0.9137 in the case of individual key-buffers and by 0.9929 in the case of shared key-buffers; again averaged over $N = 100$ trials. This means that the individual link performances are indeed sharp bounds to the true bandwidth, as even slightly overshooting (by $a = 5$ kbit/sec in out example) makes congestions highly likely. If we just look at a single evaluation of the two different methods mentioned above, we also see that the paths yielding the minimal value may differ: while working with the design of individual buffers the algorithm takes route $0 \to 1 \to 2 \to 5$ in the original graph, it does prefer $0 \to 3 \to 4 \to 5$ under the shared buffer design.

An illustration and interpretation of Proposition 1 is the following: with link performance values as given in Figure 7, a maximal flow is found at 607.54 kbit/sec. So this is the maximal traffic load that the network can handle.

Better performance is obtained when we double the size of the key-buffer in each link. Under the same set up as before, but with $L = 10$ Mbit of key-buffer and the λ-values as listed in Table 1(b), we get the blocking probabilities shown in the right column of Table 1(b). The blocking probability for an end-to-end communication in this case is $p \geq 0.8771$ when individual key-buffers are used, and $p \geq 0.9725$ when a shared buffer is employed. Finally, Proposition 1 tells that the overall end-to-end traffic from 0 to 5 is bounded above by 608.52 kbit/sec (which is the maximal flow under the respective capacities $\mu_{u\to v} = \lambda_{u\to v} - 5$ for each link).

It is important to observe that Proposition 2 explicitly is concerned with situations in which the traffic load *exceeds* the key-(re-)generation rate on the links. The converse case in which there is a positive surplus of key-material produced on each link is obviously not interesting in terms of congestion likelihoods (as the key-buffers cannot run empty in that case).

TABLE 1 BLOCKING PROBABILITIES EXAMPLES

Edge $e = u \to v$	Traffic λ [kbit/s]	Blocking prob. p_e
$0 \to 1$	305.23	0.7062
$0 \to 3$	302.32	0.7058
$1 \to 2$	304.70	0.7063
$1 \to 3$	305.35	0.7067
$1 \to 4$	305.77	0.7066
$2 \to 5$	304.28	0.7064
$3 \to 4$	305.65	0.7068
$4 \to 5$	304.22	0.7064

(a) Key-buffer size $L = 5$ Mbit

Edge $e = u \to v$	Traffic λ [kbit/s]	Blocking prob. p_e
$0 \to 1$	306.06	0.6497
$0 \to 3$	304.53	0.6493
$1 \to 2$	304.29	0.6492
$1 \to 3$	305.29	0.6495
$1 \to 4$	305.89	0.6499
$2 \to 5$	304.34	0.6494
$3 \to 4$	305.65	0.6497
$4 \to 5$	304.23	0.6493

(b) Key-buffer size $L = 10$ Mbit

IV. CONCLUSIONS

Given a quantum network, we have shown how to efficiently compute bounds to the transmission capacity and the likelihood of blocked paths due to local congestions.

A. Future Work

We focused on two specific architectures for key-buffers. Our approach and results are extensible towards more general architectures (as we considered only two “extreme” cases here) for the key-buffers as well as for the relay-regime as such (cf. [22], who propose a novel three-party quantum communication approach). It is well known that classical routing regimes face difficulties when trying to attain the upper bounds to the transmission capacity as implied by the max-flow approach (network coding is one way to resolve this dilemma). Consequently, our bounds are not necessarily tight. A closer investigation of this is subject of future work. Finally, since quantum networks have hardly reached a level of maturity beyond prototypes or lab demonstrators, reports on comparisons of our results to other competing approaches are part of future research.

B. Summary

Our analysis is entirely based on the physical topology of the network and the known key-generation rates on each link. In this work, we focused on single-path (classical) routing schemes, leaving analogous research in the field of multipath routing and network coding for future work. Our results are easy to implement with off-the-shelf algorithms, hence the proposed analysis technique is efficient in terms of computational, modeling and implementation efforts.

Despite quantum networks not having evolved beyond demonstrator prototypes yet, the possibility of setting up a high-security transmission network over existing fibre-optic lines is quite interesting. Our research here is meant as a starting point towards the construction of such infrastructures in an effective and appealing manner for the potential customer. Quality of service and service level agreements in quantum networks, unfortunately, have by now not seen the necessary attention to really bring the QKD technology to the market. Although ingenious solutions and brilliant theoretical achievements have been made, the "last mile" between lab implementation and large-scale practical business implementation needs more attention.

REFERENCES

[1] C. Bennett and G. Brassard, "Public key distribution and coin tossing," in Proceedings of IEEE International Conference on Computers Systems and Signal Processing , Bangalore, India, 1984.

[2] C. Elliott, "The DARPA Quantum Network," arXiv:quant-ph/0412029v1, 2004.

[3] M. Peev, C. Pacher, R. Alleaume, C. Barreiro, J. Bouda, W. Boxleitner, T. Debuisschert, E. Diamanti, M. Dianati, J. F. Dynes, S. Fasel, S. Fossier, M. Furst, J. D. Gautier, O. Gay, N. Gisin, P. Grangier, A. Happe, Y. Hasani, M. Hentschel, H. Hubel, G. Humer, T. Länger, M. Legre, R. Lieger, J. Lodewyck, T. Lorünser, N. Lütkenhaus, A. Marhold, T. Matyus, O. Maurhart, L. Monat, S. Nauerth, J. B. Page, A. Poppe, E. Querasser, G. Ribordy, S. Robyr, L. Salvail, A. W. Sharpe, A. J. Shields, D. Stucki, M. Suda, C. Tamas, T. Themel, R. T. Thew, Y. Thoma, A. Treiber, P. Trinkler, R. Tualle-Brouri, F. Vannel, N. Walenta, H. Weier, H. Weinfurter, I. Wimberger, Z. L. Yuan, H. Zbinden and A. Zeilinger, "The SECOQC quantum key distribution network in Vienna," New Journal of Physics, vol. 11, no. 7, p. 075001, 2009.

[4] P. Shor and J. Preskill, "Simple Proof of Security of the BB84 Quantum Key Distribution Protocol," Phys. Rev. Lett., vol. 85, pp. 441-444, 2000.

[5] R. Alleaume, F. Roueff, E. Diamanti and N. Lütkenhaus, "Topological optimization of quantum key distribution networks," New Journal of Physics, vol. 11, p. 075002, 2009.

[6] S. Rass, A. Wiegele and P. Schartner, "Building a Quantum Network: How to Optimize Security and Expenses," Springer Journal of Network and Systems Management, vol. 18, no. 3, pp. 283-299, 2010.

[7] S. Bhadra and S. Shakkottai, "Looking at Large Networks: Coding vs. Queueing," in Proceedings of the 25th IEEE International Conference on Computer Communications, Barcelona, Spain, 2006.

[8] S.-Y. R. Li, R. W. Yeung and N. Cai, "Linear Network Coding," IEEE Transactions on Information Theory, vol. 49, no. 2, pp. 371-381, 2003.

[9] R. Ahlswede, N. Cai, S.-Y. Li and R. Yeung, "Network information flow," IEEE Transactions on Information Theory, vol. 46, no. 4, pp. 1204-1216, 2000.

[10] D. H. Woldegebreal, S. Valentin and H. Karl, "Outage probability analysis of cooperative transmission protocols without and with network coding: inter-user channels based comparison," in Proceedings of the 10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems, New York, USA, 2007.

[11] S. Weber, J. G. Andrews and N. Jindal, "An overview of the transmission capacity of wireless networks," IEEE Transactions on Communications, vol. 58, no. 12, pp. 3593-3604, 2008.

[12] M. Kaynia, F. Fabbri, R. Verdone and G. Øien, "Analytical study of the outage probability of ALOHA and CSMA in bounded ad hoc networks," in Proc. European Wireless (EW), 2010.

[13] R. Vaze and R. W. Heath, "Transmission Capacity of Ad-hoc Networks with Multiple Antennas using Transmit Stream Adaptation and Interference Cancelation," http://arxiv.org/abs/0912.2630, 2009.

[14] F. Caruso, S. F. Huelga and M. B. Plenio, "Noise-Enhanced Classical and Quantum Capacities in Communication Networks," Phys. Rev. Lett., vol. 105, p. 190501, 2010.

[15] S. Watanabe, Remarks on Private and Quantum Capacities of More Capable and Less Noisy Quantum Channels, arXiv:1110.5746v1 [quant-ph], 2011.

[16] C. Le Quoc, P. Bellot and A. Demaille, "On the security of quantum networks: a proposal framework and its capacity," in Proceedings of the

International Conference on New Technologies, Mobility and Security, 2007.

[17] G. Smith, J. A. Smolin and J. Yard, "Quantum Communication with Gaussian channels of zero quantum capacity," Nature Photonics, vol. 5, pp. 624-627, 2011.

[18] G. Zhang, S. Zhou, D. Wang, G. Yan and G. Zhang, "Enhancing network transmission capacity by efficiently allocating node capability," Physical A: Statistical Mechanics and its Applications, vol. 390, no. 2, pp. 387-391, 2009.

[19] S. Weber, X. Yang, G. d. Veciana and J. Andrews, "Transmission capacity of CDMA ad-hoc networks," in IEEE Eighth International Symposium on Spread Spectrum Techniques and Applications, 2004.

[20] P. Schartner and S. Rass, "How to overcome the Trusted Node Model in Quantum Cryptography," in Proceedings of the 12th IEEE International Conference on Computational Science and Engineering, 2009.

[21] A. Mink, L. Ma, T. Nakassis, H. Xu, O. Slattery, B. Hershman and X. Tang, "A Quantum Network Manager That Supports A One-Time Pad Stream," in Proceedings of the second International Conference on Quantum, Nano and Micro Technologies, 2008.

[22] C. Le Quoc and P. Bellot, "A New Proposal for QKD Relaying Models," in Proceedings of 17th International Conference on Computer Communications and Networks, 2008.

[23] A. Abbas, "A Hybrid Protocol for Identification of a Maximal Set of Node Disjoint Paths," International Arab Journal Of Information Technology (IAJIT), vol. 6, no. 4, pp. 344-358, 2009.

AUTHORS PROFILE

Sandra König received her Bachelor and Master degree in Mathematics at the ETH Zurich, with a focus on statistics (subspecialty in regression analysis and time series analysis). Her research interests cover applications of statistics in electrical engineering as well as communication and information theory.

Stefan Rass graduated with a double master degree in mathematics (with a focus on statistics)and computer science from the Klagenfurt University in 2005, and gained a PhD degree in mathematics in 2009. His research interests include general system security, in particular applications of quantum cryptography and information-theoretic security. …

Confidential Deterministic Quantum Communication Using Three Quantum States

Piotr ZAWADZKI
Institute of Electronics
Silesian University of Technology, POLAND

***Abstract*—A secure quantum deterministic communication protocol is described. The protocol is based on transmission of quantum states from unbiased bases and exploits no entanglement. It is composed form two main components: a quantum quasi secure quantum communication supported by a suitable classical message preprocessing layer. Contrary to many others propositions, it does not require large quantum registers. A security level comparable to classic block ciphers is achieved by a specially designed, purely classic, message pre- and post-processing. However, unlike to the classic communication, no key agreement is required. The protocol is also designed in such a way, that noise in the quantum channel works in advantage to legitimate users improving the security level of the communication.**

Keywords- quantum cryptography; quantum secure direct communication; privacy amplification.

I. INTRODUCTION

The interest in quantum communication is motivated by the promise of provable security based on the laws of physics. The most mature protocols use quantum channels for secure quantum key distribution (QKD) which is further used by legitimate parties to protect communication over classic channels [1]. The content of the key resulting from QKD execution is not determined by either of users but is random and settled by the protocol completion itself. An alternative paradigm of a quantum secure direct communication (QSDC) has been investigated in the last decade [2,3]. QSDC protocols offer confidential transmission of deterministic classic information over a quantum channel without a prior key agreement. Most QSDC protocols are completely robust [4-7]. It means, that an eavesdropper cannot intercept any information without introducing errors in the transmission. Unfortunately, the absence of privacy amplification step in QSDC protocols causes that complete robustness guarantees only quasi security in perfect quantum channels – there exists finite, nonzero probability that some information is intercepted without detection. Situation is even worse in noisy environments when legitimate users tolerate some level of transmission errors. If that level is to high compared to the quality of the channel then an eavesdropper can peek some fraction of signal particles hiding himself behind acceptable QBER threshold. The possibility to intercept some part of the message without being detected renders protocol insecurity. This difficulty has been resolved by processing qubits in blocks [8-12] and/or using quantum privacy amplification [13,14]. However, such an approach requires large quantum registers which are not realizable at present with photonic techniques.

In this paper QSDC security is improved by message classic processing. The quantum protocol based on single photon transmission and not exploiting quantum entanglement is supplemented by pre- and post-processing procedures. The quantum part requires only one qubit register and such photonic quantum memory operating in high temperature has been already realized experimentally [15]. The preprocessing part is an adaptation of the transform proposed in [16] to the specific requirements of quantum communication. In the resulting protocol, contrary to many others QSDC protocols, the noise in quantum channel works in advantage for the legitimate users improving the security of communication.

II. PROTOCOL DESCRIPTION

Alice, the sender of information, is able to generate three quantum states $|0>$, $|1>$ and $|+>=(|0>+|1>)/\sqrt{2}$. Bob, the recipient of the message, is equipped with one qubit quantum register and is able to perform quantum measurements in Z and X bases. Users are connected with quantum and classic communication channels. Information in classic channel is not confidential and may be freely eavesdropped. However, it is assumed that this information cannot be modified by distrusted parties. On the contrary, the quantum channel may be tampered with no limitations – any data manipulation allowed by the laws of physics is permitted. Let $\mu = 1, \dots, N$ and $M = \{m_\mu\}$ be the data block of bits which Alice is going to send.

A. Preprocessing

1. Alice generates a random sequence of bits $K = \{k_\mu\}$. This sequence is further called a preprocessing key.
2. Alice encrypts some publicly known text T with the classic cipher of a well established reputation that produces a ciphertext $\{c_\mu\} = C = E_K(T)$ of size N using a preprocessing key K (E_K denotes encryption operation),
3. The preprocessed sequence $S = (s_\mu, s_{N+\mu})$ which will be sent is composed of two parts. The first part is formed by bitwise xoring the ciphertext from the previous step with the message bits:

$$s_\mu = m_\mu \; xor \; c_\mu, \tag{1}$$

and the second part of sequence S is calculated as

$$s_{N+\mu} = k_\mu \; xor \; s_\mu. \quad (2)$$

The encoding operation is invertible only when all bits of the sequence S are received without errors, thus some error correction code is used to protect against noise $B = ECC(S)$. Sequence B is sent to Bob via the quantum channel.

B. Communication

Qubits are processed in one-by-one manner. Alice randomly switches between control and message mode and uses classic channel to notify Bob, that the last qubit of the given data block was sent so he should proceed with the post-processing step. The sequence $\{s_{N+\mu}\}$ is sent first.

1) Control mode

1. Alice randomly prepares one of the states $|0>, |1>$ or $|+>$ which is sent to Bob.
2. Bob stores the received state in quantum register and notifies Alice.
3. Alice informs Bob that this qubit should be processed in control mode and informs Bob about the state preparation basis (Z or X).
4. Bob measures a quantum register in the basis specified by Alice.
5. If the selected basis was X, Bob knows that the measured state should be $|+>$. The appearance of $|->$ denotes a transmission error. Alice is notified about a failure.
6. If the selected basis was Z, Bob informs Alice about a measurement result and Alice compares that result with the value used in the state encoding. Bob is informed about comparison correctness.

If an error rate averaged over sufficiently large number of control qubits exceeds the correction capabilities of the ECC code then transmission is aborted before entire message is sent.

2) Message mode

1. Alice encodes bit sequence $B = \{b_\mu\}$ as states $|0>$ and $|1>$ and sends them to Bob.
2. Bob stores the received state in quantum register and notifies Alice.
3. Alice informs Bob that this qubit should be processed in message mode.
4. Bob measures quantum register in Z basis, stores the measurement result as b'_μ in classic memory and notifies Alice that he is ready for the reception of the next qubit.

C. Postprocessing

ECC data is used to correct errors on the received sequence $S' = ECC^{-1}(B')$,

1. The preprocessing key is recovered as $k'_\mu = s'_\mu \; xor \; s'_{N+\mu}$.
2. The ciphertext sequence is again calculated as $C' = E_{K'}(T)$ and the message decoded as $m'_\mu = s'_\mu \; xor \; c'_\mu$.

If any of the bits in the sequence S' is incorrect then the key K' is also incorrect and the sequence C' is completely different from C (this behavior is guaranteed by the properties of the classic cipher). It follows that Eve can recover a message only when she correctly intercepts entire sequence S. Incorrect detection on only one position results in (almost) random decoded message.

III. Analysis

An important step in studying protocol security is an investigation of its robustness. Robustness of the protocol informs how large disturbance is introduced by an eavesdropper intercepting some information. QKD protocols can be secure when they are partly or completely robust [4]. However, security requirements for QSDC protocols are much more stringent because of the absence of privacy amplification step. As a result partly robust QSDC protocols are considered insecure and complete robustness guarantees only quasi security, i.e. there is a finite, non-zero probability that eavesdropper intercepts some part of the message without being detected. Although similar property also holds for classic cipher, the problem with QSDC security lies in fact, that offered eavesdropping detection probability is relatively low and Eve is detected with reasonable probability only after sufficiently large number of protocol cycles. The pre- and post-processing steps introduced herein provide all or nothing logic in the message interception possibility. Thus proposed protocol is insecure only when its quantum part is not robust. Contrary, if the quantum part is partly robust or completely robust then introduced classic pre- and post-processing steps provide protocol computational security. In the following it will be proven that quantum part is robust in lossless quantum channels and partly robust in a lossy case. This renders that Eve intercepts no useful information and protocol is secure. The provided security margin is related to the quantum transmission quality and QBER introduced by the eavesdropping.

Let us consider robustness of the protocol in the noiseless quantum channel case. As it follows from the Stinespring's dilation theorem, the most general quantum operation, which may be performed on the signal qubit by an eavesdropping Eve is described by an unitary operation that entangles the signal qubit with the ancilla system of dimension 2^2. Such a transformation is described by four complex numbers α_k, β_k

$$U|0>|\varphi> = \alpha_0 |0>|\varphi_{00}> + \beta_0 |1>|\varphi_{01}>, \quad (3)$$

$$U|1>|\varphi> = \alpha_1 |0>|\varphi_{10}> + \beta_1 |1>|\varphi_{11}>. \quad (4)$$

where $|\varphi_{kl}>$ denotes Eve's probe states and $|\varphi>$ is the initial state of the ancilla which is not entangled with the signal qubit. The normalization ensures that $|\alpha_k|^2 + |\beta_k|^2 = 1$. The same entangling operation has to be applied to the message and control qubits because Eve acquires information what mode has been used after the qubit has been stored in Bob's register. Eve is detected with probability $|\beta_0|^2$ and $|\alpha_1|^2$ when legitimate users test in Z basis. It follows from (3) and (4) that

$$U|+>|\varphi>=|+>(\alpha_0|0>|\varphi_{00}>+\beta_0|1>|\varphi_{01}>)+$$
$$+|+>(\alpha_1|0>|\varphi_{10}>+\beta_1|1>|\varphi_{11}>)+$$
$$+|->(\alpha_0|0>|\varphi_{00}>-\beta_0|1>|\varphi_{01}>)+$$
$$+|->(\alpha_1|0>|\varphi_{10}>-\beta_1|1>|\varphi_{11}>), \quad (5)$$

where $|->=(|0>-|1>)/\sqrt{2}$. Thus she is detected in X basis with probability

$$(|\alpha_0|^2+|\beta_0|^2+|\alpha_1|^2+|\beta_1|^2)/4=1/2 \quad (6)$$

because U is unitary. Thus overall Eve's detectability is minimized when $|\beta_0|^2=|\alpha_1|^2=0$. But in that case Eve's probe space is limited to two states and an entangling operation may be reduced to simple CNOT in which signal qubit is used as the control one. At the same time such operation provides maximal information about the state of the signal qubit when Eve performs measurements in Z basis. The quantum part of the protocol is completely robust because Eve can't intercept any information without risking to be detected. However, robustness of the quantum transmission implies only quasi security of the QSDC protocol.

Let us further consider how introduced preprocessing improves the security characteristic of the protocol and assume that legitimate users use ECC code able to recover from a given QBER although they operate in noiseless channel. Such assumption is favorable to the eavesdropper, as she can now intercept some signal particles and her actions will be undistinguishable from the noise. Because Eve is detected in X with probability $1/2$, it means that she can peek $2N \times QBER$ of signal qubits per block without inducing an alarm. This is the best case scenario for the eavesdropper. If the channel has been noisy Eve would have to attack a less percentage of particles to hide herself behind the total limit of errors. Thus it may be assumed that she knows $2 \times QBER$ fraction of s_μ and $s_{N+\mu}$ sequences and the rest part of these sequences remains random to her.

The preprocessing key is recovered as $k'_\mu = s'_\mu \; xor \; s'_{N+\mu}$. But correctly recovered are only these fragments for which bits on corresponding positions of sequences s'_μ and $s'_{N+\mu}$ are correct and that happens with probability $(2\,QBER)^2$. Thus to recover the message $m_\mu = s_\mu \; xor \; \mathrm{E_K(T)}$ malicious Eve has to attack a key space $[1-(2\,QBER)^2] \times N$ of a well established cipher and guess $[1-(2\,QBER)] \times N$ bits of the sequence s'_μ. It follows that number of bits that have to be tested in brute force attack exceeds N for $QBER \leq (\sqrt{5}-1)/4 \cong 0.31$. Available presently quantum channels may provide a better performance. Thus computational complexity of an attack on the protocol exceeds complexity of brute force guessing of the message contents. Protocols with such property are regarded in classic cryptography as secure. It is worth noting that although computational complexity of the brute force attack has been considered above, the proposed protocol does not require establishment of the shared secret for secure operation. Moreover, the block cipher used in the protocol works only in encryption mode, thus may be replaced by another cryptographic primitive providing randomization of the input data, for instance, a stream cipher generator or keyed hash function.

The performance of the described protocol is determined by the overhead related to the transmission of test qubits and a check block $s_{N+\mu}$. The number of test qubits must provide reliable estimation of the channel quality within one data block because decision about channel reliability should be taken before the sequence s_μ with encoded message is sent. As a matter of fact the control protocol cycles may be disabled during encoded message transmission as at this point of protocol execution is too late for the eavesdropping detection. The overhead related to transmission of the check sequence may be diminished for messages longer than N bits. In such case message is padded and divided onto blocks $m_\mu^{(0)}, m_\mu^{(1)}, \ldots, m_\mu^{(M)}$ and each block is processed and sent independently $s_\mu^{(k)} = m_\mu^{(k)} \; xor \; c_\mu$. The block with encoded preprocessing key $k_\mu \; xor \; s_\mu^{(0)} \; xor \ldots xor \; s_\mu^{(M)}$ is sent as the first one.

IV. Conclusion

A single photon based protocol for quantum secure direct communication has been proposed. Although its quantum part is only quasi secure, the classic pre- and post-processing of the message improves protocol security to the desired level. The introduced protocol has very small demands on quantum resources and can be, in principle, practically implemented in the near future. Although the protocol is not unconditionally secure, the provided security margin is high in noisy quantum channels. It also offers some advantages compared to quantum key agreement schemes proposed so far. The unconditional security of QKD protocols has been proved in the limit of the infinite length of the block being processed and the length of the secret key is less than 50% of qubits sent. However, efficiency of QKD scales very badly with the decrementation of the sequence size and for moderate blocks of size 10^6 the efficiency does not exceed 2% [17]. The proposed approach offers improved efficiency at the price of the computational security. It is also more versatile as it may be used for confidential exchange of short sensitive messages without key agreement and for regular QKD as well. Protocol also offers also some advantages compared to presented so far QSDC schemes [2,5] – the quasi security limitation has been conquered without requirement of large quantum memory registers which are out of the reach of the present state of the art technology.

References

[1] N. Gisin, G. Ribordy, W. Tittel and H. Zbinden, "Quantum cryptography", Rev. Mod. Phys., vol. 74, pp. 145-195, 2002

[2] K. Boström and T. Felbinger, "Deterministic secure direct communication using entanglement", Phys. Rev. Lett., vol. 89, pp. 187902, 2002

[3] A. Beige, B. G. Englert, C. Kurtsiefer and H. Weinfurter, "Secure communication with a publicly known key", Act. Phys. Pol., vol. 101, pp. 357-368, 2002

[4] M. Boyer, R. Gelles, D. Kenigsberg and T. Mor, "Semiquantum key distribution", Phys. Rev. A, vol. 79, pp. 032341, 2009

[5] M. Lucamarini and S. Mancini, "Secure deterministic communication without entanglement, Phys. Rev. Lett., vol. 94, pp. 140501, 2005

[6] K. Boström and T. Felbinger, "On the security of the ping-pong protocol", Phys. Lett. A, vol. 372, pp. 3953-3956, 2008

[7] G. L. Long, F. G. Deng, C. Wang, X. H. Li, K. Wen and W. Y. Wang, "Quantum secure direct communication and deterministic secure quantum communication", Front. Phys. China, vol. 2, pp. 251-272, 2007

[8] F. G. Deng, G. L. Long and X. S. Liu, "Two-step quantum direct communication protocol using the Einstein-Podolsky-Rosen pair block", Phys. Rev. A, vol. 68, pp. 042317, 2003

[9] S. Lin, Q. Y. Wen, F. Gao and F. C. Zhu, "Quantum secure direct communication with □-type entangled states", Phys. Rev. A, vol. 78, pp. 064304, 2008

[10] J. Wang, Q. Zhang and C. Tang, "Quantum secure direct communication without a pre-established secure quantum channel", Int. J. Quant. Inf., vol. 4, pp. 925–934, 2006

[11] G. Gao, "Two quantum dialogue protocols without information leakage", Opt. Commun., vol. 283, pp. 2288-2293, 2010

[12] G. F. Shi, X. Q. Xi, M. L. Hu and R. H. Yue, "Quantum secure dialogue by using single photons", Opt. Commun., vol. 283, pp. 1984-1986, 2010

[13] F. G. Deng and G. L. Long, "Reply to 'Comment on "Secure direct communication with a quantum one-time-pad"'", Phys. Rev. A, vol. 72, pp. 016302, 2005

[14] D. Fu-Guo and L. Gui-Lu, „Quantum privacy amplification for a sequence of single qubits", Commun. Theor. Phys., vol. 46, pp. 443, 2006

[15] K. F. Reim, P. Michelberger, K. C. Lee, J. Nunn, N. K. Langford and I. A. Walmsley, "Single-photon-level quantum memory at room temperature", Phys. Rev. Lett., vol. 107, pp. 053603, 2011

[16] R. L. Rivest, "All-or-nothing encryption and the package transform", LNCS, vol. 1297, pp. 210-218, 1997

[17] V. Scarani, H. Bechmann-Pascanucci, N. J. Cerf, M. Dusek, N. Lutkenhaus, M. Peev, "The security of practical quantum key distribution", Phys. Rev. Lett., vol. 107, pp. 053603, 2011

AUTHORS PROFILE

P. Zawadzki received M.S. degree in theoretical physics from Silesian University, Katowice, Poland, in 1989 and Ph.D. degree in electromagnetic engineering from Silesian University of Technology, Gliwice, Poland, in 1998. His research interests include numerical simulation of interactions between telecommunication infrastructure with the lightning electromagnetic pulse, data protection in telecommunication networks and quantum information processing in the context of quantum cryptography.

A Novel Implementation of RISI Controller Employing Adaptive Clock Gating Technique

M.Kamaraju
Professor & Head, Dept.of ECE
Gudlavalleru Engineering College, Gudlavalleru, INDIA

Praveen V N Desu
M.Tech ES group, Dept of ECE
Gudlavalleru Engineering College, Gudlavalleru, INDIA

***Abstract*—With the scaling of technology and the need for higher performance and more functionality power dissipation is becoming a major issue for controller design. Interrupt based programming is widely used for interfacing a processor with peripherals. The proposed architecture implements a mechanism which combines interrupt controller and RIS (Reduced Instruction Set) CPU (Central processing unit) on a single die. RISI Controller takes only one cycle for both interrupt request generation and acknowledgement. The architecture have a dynamic control unit which consists of a program flow controller, interrupt controller and I/O controller. Adaptive clock gating technique is used to reduce power consumption in the dynamic control unit. The controller consumes a power of 174µw@1MHz and is implemented in verilog HDL using Xilinx platform**

Keywords- Interrupt; Controller; Clock gating; power.

I. INTRODUCTION

The Interrupt Controller [1-2] is a device commonly found in computer systems (both single-processor and multiprocessors) which deals with interrupts generated by the peripherals and the processors handle the interrupt priorities, and delegates the execution to a processor.

The general purpose processors provide one or more interrupt request pins that allows external devices to request the service provide by CPU. Consider a case in which processor can handle a large number of interrupts which are come from external devices. The design requires a separate interrupt controller which is interfaced to the processor. This increases the complexity of design. More over the processor needs some extra- interfacing Circuits which decreases the performance and increase the power consumption of the overall system. The proposed architecture combines the interrupt controller and RIS CPU employs an adaptive clock gating to reduce the overall power consumption.

Power [18] is the one of the design constraint, which is not only applied to portable computers and mobile communication devices but also for high-end systems. Power dissipation becomes a bottleneck for future technologies.

In the early days designers treat the clock signal should not be disabled or disturbed. But clock signal is a major source for power dissipation and it is a dynamic in nature because clock signal is feed into several blocks in the processor. Because all the blocks usage varies within and across a processor, all the blocks not used all the time and gives a chance to reduce the power consumption of unused blocks. Clock gating is an efficient technique to reduce the dynamic power dissipation. By anding the clock signal with gated control signal clock gating technique disables the clock signal to the block when the block is unused. Adaptive clock gating technique [4-6] is one of thetechnique used to reduce the dynamic power of the clock. In this technique clock gating enable signal generated by the block itself depending upon the usage and this technique will reduce the burden on the control unit for generating clock gating signal.

Interrupt handling mechanism [8] provides how the interrupt is handled by processor. There are various clock gating techniques [4][5] to reduce the dynamic clock power dissipation. The interrupt controller [1][2] takes two cycles (one cycle for generating the interrupt request and another cycle for Acknowledgement) to process the interrupt. RISI Controller takes only one cycle for both interrupt request generation and acknowledgement. The following section provides a brief overview of architecture of the RISI CONTROLLER and explanation about the implementation and hardware consideration. Along with a brief description about each block present in the architecture is given. Finally a few notes on simulation results

II. ARCHITECTURE OF RISI CONTROLLER

The reduced instruction set interrupt controller (RISI Controller) architecture mainly consists of ALU (Arithmetic and logical Unit), Port Controller, Interrupt controller and Register Array and its block diagram is shown in the Fig. 1. It contains RISCPU, Interrupt controller, Port controller and Program flow controller. These blocks are connected by internal buses.

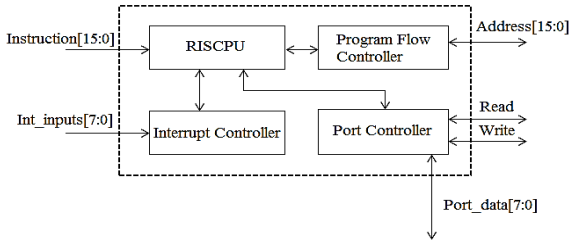


Figure 1. Block diagram of a RISI Controller

The internal architecture of RISI Controller is shown in the figure2.

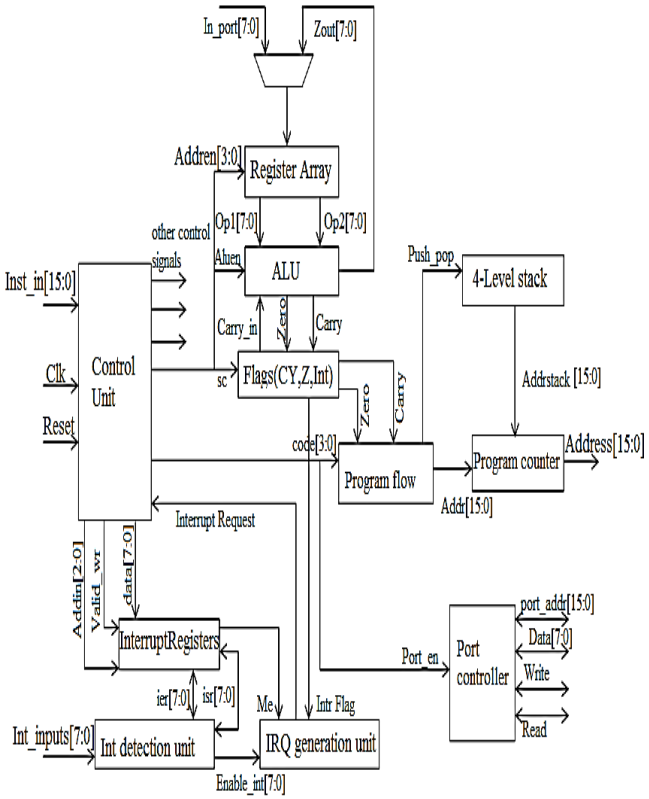


Figure 2.Internal architecture of RISI CONTROLLER

The instruction length of RISI Controller is 16-bit wide.RISI Controller has three flags namely carry, zero and interrupt flags. Both zero and carry flags are affected only during the execution of arithmetic and logical instructions and these are also useful for determine the flow of execution when branch and jump instructions take place. CPU checks the interrupt flag after completion of every instruction to know whether interrupt is available or not.ALU is capable of performing the Arithmetic (Like Addition and subtraction) and Logical operations (like And, Or, Xor and Cmpl). There are no special purpose registers in the CPU like accumulator and there is no priority among them.

RISI Controller has multi read port and single write port. Generally read operations are performing during the positive edge of the clock and write operation is performing during negative edge of the clock. Stack is used to store up to four addresses during interrupt and Branch related instructions. Port controllers take care of the read and write operation. An 8-bit address value provided on the PORT bus together with a READ or WRITE strobe signal indicates the accessed port. The port address is either supplied in the program as an absolute value or specified indirectly as the contents of any of the eight registers

There are some specific instructions useful for the controlling of interrupt controller present in the RISI CONTROLLER.

III. INTERRUPT CONTROLLER

Modern CPU's [15, 16] provide one or more interrupt request pins that allows external devices to request the service provide by CPU. Interrupt controller are used to increase the number of interrupt inputs available to CPU. The block diagram, of interrupt controller is shown in figure3.

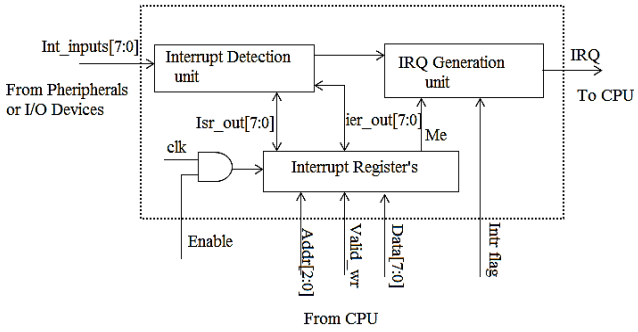


Figure3.Block diagram of Interrupt Controller

Interrupt controller composed with three blocks. They are Interrupt Register block, Edge interrupt detection unit and Interruptrequest generation unit. Interrupts are identified by interrupt detection unit during the negative clock edge of the clock. Whenever interrupts are detected, check for the corresponding interrupt input masked or not. Unmasked interrupt input set the corresponding bit in the interrupt status register. IRQ generation unit generates the interrupt request by using the IVR contents. Interrupt request reaches the CPU send an acknowledgement signal.

Int_inputs are used to monitor the interrupts coming from various peripherals or external devices. Each interrupt register has a unique address and identified by using *Addr* input. To write the contents of *Data* input into the interrupt registers require a high *valid_wr* input. *Intr flag input* indicates the status of interrupt flag present in the CPU.

Interrupt Detection Unit detects the interrupts coming from peripheral or external devices and actives the logic to generate enable interrupt to controller. It monitors the interrupt inputs composed of interrupt signal coming from external devices or peripherals and rises enabled interrupts according to arrival signals, Interrupt Request (IRQ) Generation unit contains the Generation logic of the Interrupts towards the processor.Interrupt requests generation is also configurable as either a pulse output for an edge sensitive request or as a level output that is cleared when the interrupt is acknowledged. Interrupt Registers handles the interrupt priorities, deciding which, interrupts are enabled or disabled and managing of interrupt acknowledgements. It contains the following Registers

Interrupt Status Register (ISR) indicates which interrupts are active and the format is shown in the figure 4. All bits in the ISR are set to zero default. Any bits are set to '1' indicates that the corresponding interrupt is active otherwise no active interrupts are available

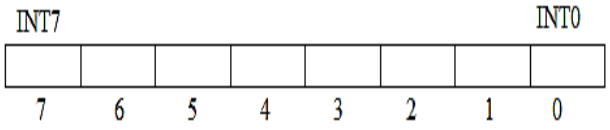


Figure 4. Interrupt Status Register

Interrupt Pending Register (IPR) gives the information about the interrupts that are both active and enabled. By default all the bits in IPR are set to zero. Any bit set to '1' indicates that the corresponding interrupt is waiting for processing and '0' indicate no interrupt isavailable. The IPR is shown in the figure 5

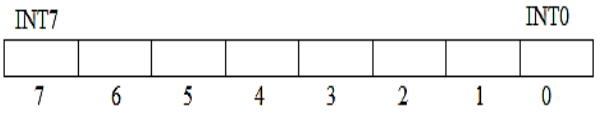


Figure 5. Interrupt Pending Register

Interrupt Enable register (IER) keeps track which interrupts are allowed to be handled. Writing a '1' to a bit in this register enable the corresponding interrupt input signal. Writing a '0' to a bit disable or mask the corresponding interrupt input signal

Interrupt Acknowledge Register (IAR)is used to disabling the interrupt request with selected interrupt input. Writing a '1' to a bit location will clear the interrupt request that was generated by the corresponding interrupt input and Writing '0' does nothing. The IAR is shown in the figure 6

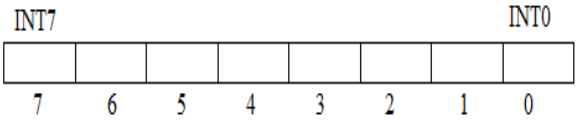


Figure 6. Interrupt Acknowledge Register

Interrupt Vector Register (IVR) contains the ordinal value of the highest priority, enabled, active interrupt input. INT0 (always the LSB)is the highest priority interrupt inputand each successive input to the left has a correspondingly lower interrupt priority. If no interrupt inputs are active then the IVR will contain all ones. The Interrupt Vector Register (IVR) is shown in the figure 7.

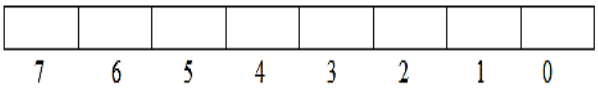


Figure 7.Interrupt Vector Register

Master Enable Register (MER) is used to enabling the interrupt requests to the processor. Writing a '1' to ME bit enables the IRQ output signal and '0' to ME bit disable the IRQ output signal in other words masking all the interrupt inputs. The Master Enable Register (MER) is shown in the figure 8.

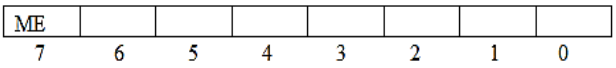


Figure 8.Master Enable Register

The following are the important features of Interrupt controller present in the RISIController. They are i) Priority between interrupt requests is determined by vector position. The least Significant Bit (LSB, bit 0) has the highest priority. ii) Interrupt enable register for selectively disabling or enabling of individual interrupt inputs. iii) Master enables register for disabling interrupt request. iv) Easily cascaded to provide additional interrupt inputs. v) Low power and less area

IV. IMPLEMENTATION

The RISI CONTROLLER is implemented using Xilinx platform on Virtex4 FPGA Family in VerilogHDL. The flowchart for the Interrupt controller is shown in the figure 9. Application specific instructions [13] were designed for controlling the interrupt controller along with general purpose instruction set.

By using Application specific instructions [13] processor can perform several operations on interrupt controller. The operations like masking or unmasking of interrupts disable interrupt request for the execution of important instructions. Interrupt controller uses fixed priority algorithm for generating the interrupt request.

The flow chart for the interrupt controller is shown in the figure 9 and is explained below. Various peripherals or an external device wants the services provided by CPU. They generate an interrupts to interrupt controller. Interrupt coming from various external devices are identified by edge interrupt detection unit during the negative half cycle of the clock. Edge detection unit in the interrupt controller check whether the interrupts are masked or not. This information is present in the IER register.

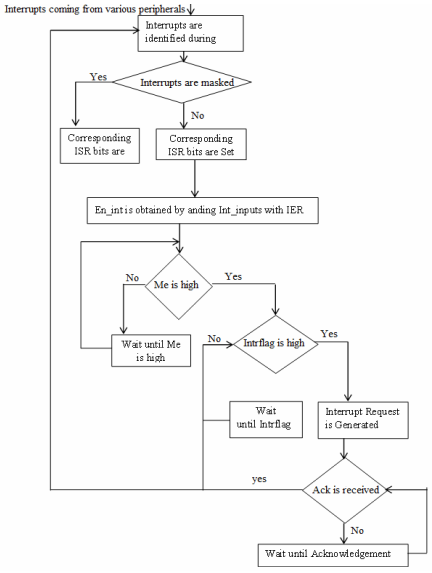


Figure 9. Flow chart for interrupt controller

Enable interrupts are obtained by anding the Interrupt_inputs with the contents of IER register. This information is stored in the ISR register.

The MER must be programmed based on the intended use of the Interrupt Controller.There are two bits in the MER: the Hardware Interrupt Enable (HIE) and the Master IRQ Enable (ME). The ME bit must be set to enable the interrupt request output.Check whether the Msb bit (ME) in MER register is set or not. If ME bit in the MER register is not set the interrupt requests are not generated. Check the status of the interrupt flag. If the flag bit is set Interrupt request is generated and wait for acknowledgement which is coming from the processor. Next interrupt request is generated only when the acknowledgement is received otherwise; wait until the acknowledgement is received.

By the execution of the following code the RISI Controller disables the interrupt request and enables it again after some time. The simulation results regarding the execution of the program1 are shown in the results section

Program 1

```
MID 00          -----  Disables the interrupt Request
MOV   R4, 5F ----- Load the immediate data5F into register R4
MOV   R5, 5F -----  Load the immediate data 5F into register R5
ADD    R4, R5  -----Add the contents of R5with R4 and result is
                    Stored in R4
MIE 80          ----- Enables the interruptrequest
```

RISIController is capable to handle 8 interrupt inputs at a time. Interrupt controller uses positive clock cycle for interrupt request generation whereas negative clock cycle used for receiving an interrupt Acknowledgement which is coming from CPU.

V. RESULTS

All the modules of RISI CONTROLLER are simulated and verified using the Xilinx tools. The simulation result of Interrupt controller shown in figure 10. When the reset is high all the register present in the interrupt controller are loaded with default values. Interrupts are identified during the negative edge of the clock signal. The content of IER indicates whether the interrupts are mask or unmask. En_int is obtained by anding the content of IER with Int_inputs.

Each register present in the interrupt controller have a unique address. Registers are identified by using *addr* input. The content of*En_int* is also stored in the ISR. Interrupt request is generated only if both the *ME* and Intr flag arehigh. Priority between the interrupts is generated by using the fixed priority scheme. The contents of *Clr_isr* are set at the time of interrupt request generation.

Top level timing diagram between interrupt controller and RISCPU is shown in figure 11. Whenever the CPU receives interrupt request, wait for the completion of current instruction and store the address of next instruction in the stack. Load the contents of program counter with address FF and send an acknowledgement to the Interrupt Controller by activating the *addrin* and *valid_wr* output.

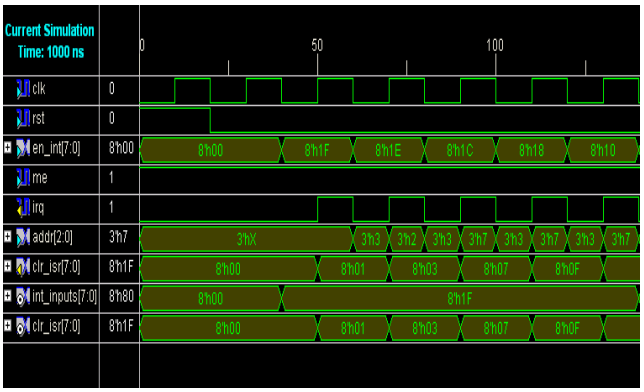


Figure 10.Top level timing diagram of Interrupt Controller

Status of *intr flag* is changed from high to low when the interrupt request is received. CPU does not receive any interrupt even though *Int_inputs* are active because *Me* output is low indicate that interrupt request is disabled. The status of *Me*output is high then only the CPU receives the interrupt request and it is shown in the figure 11.

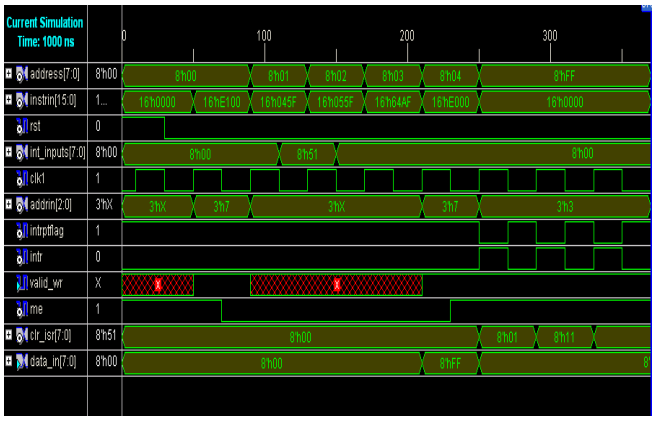


Figure 11: Top level timing Diagram between the Interrupt controller and RIS CPU

The top level module with input and out signals of the RISI Controller and Interrupt Controller is shown in figures 12, 13 and the RTL schematic of the Interrupt Controller is shown in the figure 14.

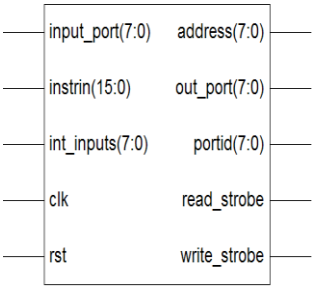


Figure 12. Signals of RISI CONTROLLER

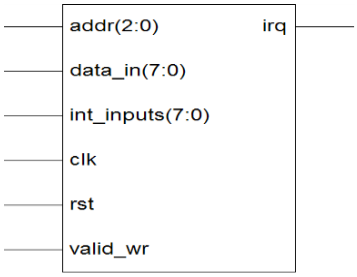


Figure 13. Signals of Interrupt Controller(IC)

The RTL schematic of the Interrupt Controller of RISI Controller is shown in figure 14. The Interrupt controller consists of a Register block, edge interrupt detection unit and IRQ generation unit. Interconnection between the blocks and the input output signals of the Interrupt Controller module are also shown in the figure 14.

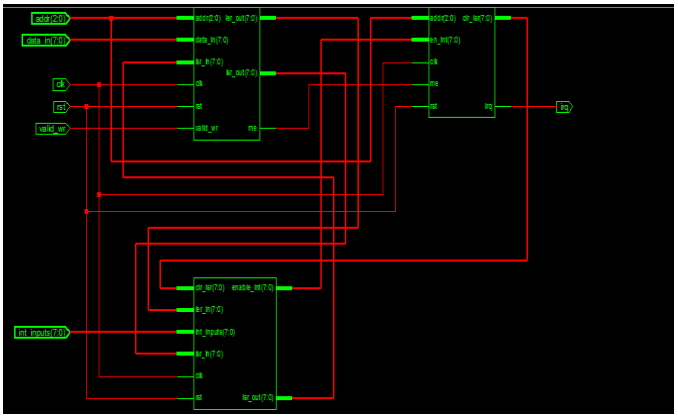


Figure 14.RTL Schematic of Interrupt controller of RISI Controller

Figures 15 and 16 represents the layout of the RISI Controller and Interrupt controller implemented onto the Vertex4 FPGA family. In the Fig. 15 and 16 the colored area represents the components of the RISI CONTROLLER that are placed on the FPGA and the place and routed diagrams of the Interrupt Controller are shown in figures 17 and 18 respectively

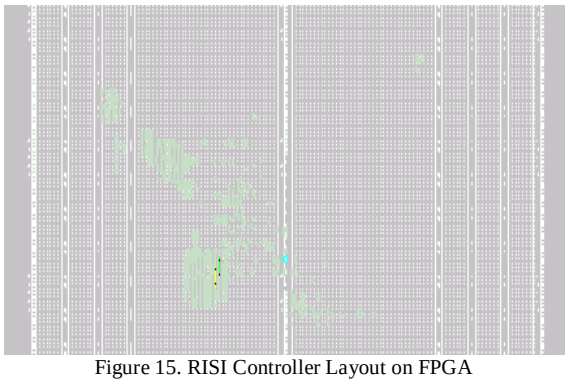


Figure 15. RISI Controller Layout on FPGA

Placement involves deciding where to place all electronic components, circuitry, and logic elements in limited amount of space. This is followed by routing, which decides the exact design of all the wires needed to connect the placed components.

Figure 16.Interrupt controller(IC) layout on FPGA

This must implement all the desired connections by following the rules and limitations .The complete placed and routed diagram of the RISI Controller is shown in figure 17.

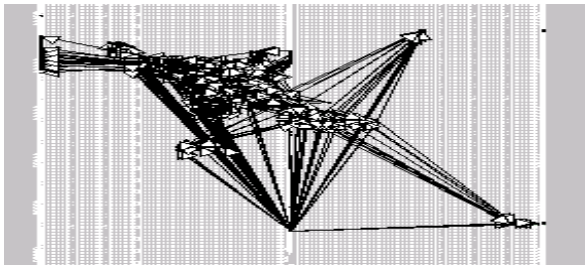


Figure 17.Place and Route diagram of RISI CONTROLLER

The place and route diagram of theInterrupt controllerof RISI Controller is shown in figure 18.

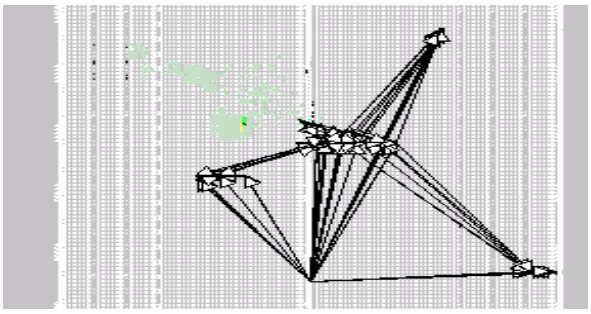


Figure 18. Place and Route diagram of Interrupt Controller(IC)

In the Table I the comparison of different parameters related to the area of RISI Controller and Mpcore were specified. The graph clearly shows that the number of slices and LUTs utilized by the RISI Controller less than that of the Mpcore. The graph also indicates in othersense that the area occupied by RISI Controller on the FPGA is less.

TABLE I.COMPARISON OF DEVICE UTILIZATION PARAMETERS OF MPCORE AND RISI CONTROLLER

Parameter	Mpcore	RISI CONTROLLER
Number of slice Registers	349	318
Slices containing related logic	496	467
Gate count	7360	7162

The comparison of different parameters related to RISI Controller signal delays were shown in the figure 19. The Net Skew for clock is the difference between the minimum and maximum routing only delays for the clock.

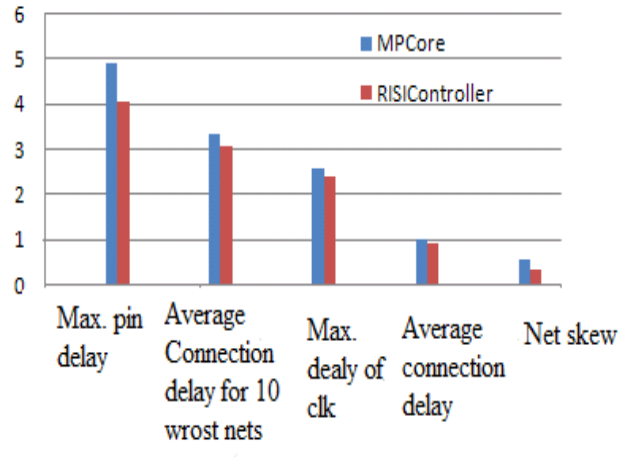


Figure 19.Comparison of delay parameters for Mpcore and RISI Controller

Power is one of the main constrain in the design of controller. Clock signal [4] contribute major power consumption source. To reduce the clock power dissipationclock gating techniques are employed. The power consumption of the RISI CONTROLLER was described in the figures 20 and 21. The figures also give us a comparative view of the power consumption of the Mpcore and RISI CONTROLLER cores.

In the figure 21 the deviation of the curves indicate that the operating voltage increases the power consumption of the Mpcore increases more rapidly when compared with designed controller power. Similarly figure 20 describes the variation of power with respect to frequency for the Mpcore and RISI CONTROLLER.

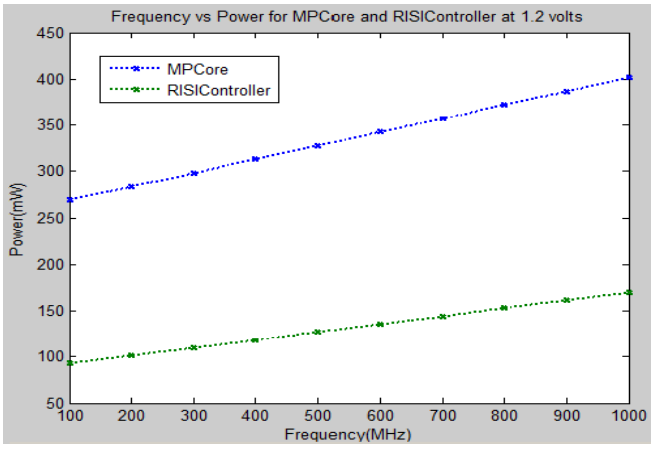


Figure 20. Power consumption of RISI CONTROLLER at 1.2v

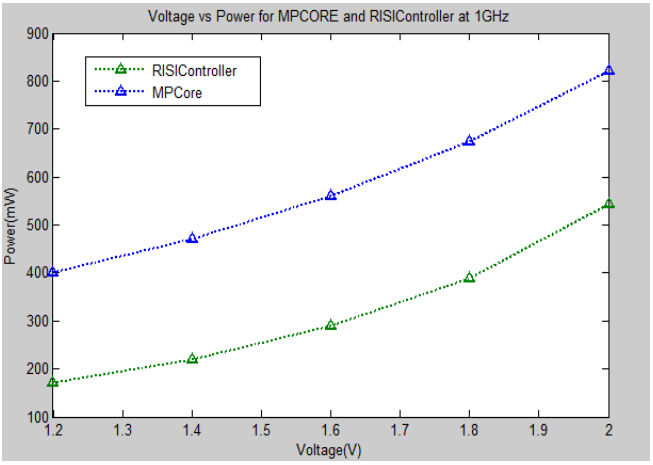


Figure 21.Plot b/w Voltage and Power of Mpcore and RISI Controller at 1 GHz

The Power-delay product is simply the product of the power consumption and the time delay. The smaller value of the power-delay product, performance of the design is better. Since this RISI CONTROLLER has almost negligible power-delay product value, it indeed has a better performance in terms of the speed and power dissipation.

TABLE II.COMPARISON OF MPCORE AND RISI CONTROLLER

PARAMETER	Mpcore	RISI Controller
Total number of gates	7,360	7,162
Memory usage for synthesis	247.5 MB	234.3 MB
Max frequency	0.928GHz	1.08GHz
Power consumption(@1.2V,1GHz)	401.53mW	170.39mW
Power Delay Product	1027.9pJ	142.61pJ

Table II shows the comparison of different parameters related to the Mpcore and RISI CONTROLLER; it has better performance than the Mpcore.

VI. CONCLUSION

As the amount of data transferred between the main processing unit and peripheral devices increases, the frequency of interrupts from peripheral devices also increases. The RISI controller designed integrates a RISCPU and Interrupt controller onto the single chip. The advantage of the designed chip is, it can handle an interrupt fast and effectively. It occupies less area and consumes less power. More over an integrated CPU in the design performs the necessary operations related to interrupt controller apart from the regular operation. The scope for increasing the number of interrupts up to databus width isprovided in the design and also extended to multiprocessor.

REFERENCES

[1] A. Tumeo, M. Branca, L. Camerini, M. Monchiero,G. Palermo,F. Ferrandi and D. Sciuto, "An Interrupt Controller for FPGA-based Multiprocessors", International Conference on Embedded Computer Systems: Architures, Modeling and Simulation, July 2007,pp. 82-87.

[2] Wei Chipni, Li Ahao lin, Zheng Qingwei,Ye Jianfei, and Li Shenglong, "Design of a configurable multichannel interrupt controller",Second Pacific-Asia Conference on Circuits,Communications and System, vol. 1,Aug 2010, pp. 327–330.

[3] Pizhou Ye,and Chaodong Ling, "A RISC CPU IP core",Second International Conference on Anti-counterfeiting,Security and Identificatiom ,August(2008), pp. 356–359,2008.

[4] H. Jacobson,P. Bose, Zhigang Hu,A. Buyuktosunoglu, V.Zyuban,R. Eickemeyer,L. Eisen,J. Griswell,D. Logan,Balaram Sinharoy, and J. Tendler," Strenching the limits of clock-gating efficiency in server-class processors",Eleventh International Symposium on High-performance Computer Architecture,Feburary (2005), pp . 238- 242,2005.

[5] Xiaotao Chang, Mingming Zhang, Ge Zhang, Zhimni Zhang and Jim Wang,"Adaptive Clock Gating Technique for Low Power IP Core in SoC Design",IEEE International Symposium on Circuits And Systems,May(2007),pp. 2120 – 2123, 2007..

[6] S. Ghosh, D. Mohapatra, G. Karakonstantis and K. Roy," Voltage Scalable High-speed Robust Hybrid Artihmetic Units Using Adative Clocking"IEEE Transactions on Very Large Scale Integration Systems,September(2010),Vol. 18, pp.1301- 1309,2010.

[7] Hai Li, S. Bhunia,Y. Chen,T.N. Vijaykumar and K.Roy," Deterministic clock gating for microprocessor power reduction", Ninth International Sympossium on High-Performance Computer Architecture, February (2003), pp. 113 – 122, 2003.

[8] E. Ozer, S.W. Sathaye, K.N. Menezes, S. Banerjia, M.D. Jenningsand T.m. Conte," A fast interrupt handling scheme for VLIW processors",International Conference on Parallel Architecures and Compilation Techniques,October(1998),pp. 136 -141 ,1998.

[9] G. Kane and J. Heinrich," MIPS RISC Architecture: reference for the R2000,R3000,R6000 and the new R4000 Instruction set computer Architecture", Prentice-Hall,Englewood Cliffs,NJ,1992.

[10] Krall. A, " An extended Prolog instruction set for RISC processors" in VLSI for Artifical Intelligence and Neural Networks,J.G. Delgado-Frias and W.R. Moore, Eds-Plenum Press, New York,pp. 101-108,1991.

[11] Motorola Inc."M6800 8/16/32- Bit MICROPROCESSOR",avilable athttp://www.freescale.com/files/32bit/doc/ref_manual/MC68000UM.pdf

[12] J. Rose,A.E. Gamaland A. Sangiovanni and A. Vincentelli," Architecture of Field-Programmable Gate Arrays", Proc. IEEE,Vol81,no 7,July,pp. 1013-1020,1993.

[13] J. Vanprate, Gossens .G, D. Lanner, and H. De man," Instruction set definition and insstruction set selection",in Proc. Of seventh International Symposium on High-Level Synthesis,pp. 11- 16,1994.

[14] ARM11 Mpcore available athttp://www.arm.com.

[15] IBM. Multiprocessor Interrupt Controller Data Book,March,2006.

[16] Xilinx OPB Interrupt Controller (v1.00c),January 2005.

[17] A. De Gloria,Paolo Faraboschi and Mauro Olivieri," A Self TimedInterrupt Controller: ACase study in Asynchronous Micro_Architecture Design " Seventh Annual IEEE International ASIC Conference and Exhibit ,pp. 296 – 299,1994.

[18] G. Palumbo,F. Pappalardo,and S. Sannella " Evaluation on power reduction applying gated clok approes",IEEE International Symposium on Circuita and Systems,Vol.4,pp.IV-85-IV-88,2002.

[19] Nakasimha .K,S. Kusakabe,H. Taniguchi,and M. Amamiya " Design and implementaion of Interrupt packaging mechanisum",International Workshop on Innovative Architecture for Future Generation High-performance processors and systems", pp. 95-102,2002

[20] J.E. Smith, and A.R. Pleszkun," Implementing Precise Interrupts in Pipelined Processors" IEEE Trans. On Comp.,Vol.37,pp. 291-299,1975.

[21] Horelick Dale " simple Versatile Camac Crate Controller and Interrupt Priority Encoding Module ", IEEE Transaction on Nuclear Science,Vol. 22,pp. 517- 520,1975.

[22] Qiurong Wang " An Interrupt Management Scheme Based on Application in Embedded system",Multimedia and Informetion Technology,pp. 449-452,2008.

[23] Kamaraju.M, Lal Kishore.K, Tilak.A.V.N "A Novel Implementation of Application Specific Instruction-set Processor (ASIP) using Verilog " WASET Issue 59, NOV 2011.

Strength of Quick Response Barcodes and Design of Secure Data Sharing System

Sona Kaushik
Student of Masters in Technology,
Birla Institute of Technology,
Mesra, Ranchi, India

***Abstract* - With the vast introduction of the wireless world, the exchanged information now is more prone to security attacks than ever. Barcodes are the information careers in the form of an image. Their various applications have been discussed in brief and also the structure, symbology and properties of barcodes. This paper aims to provide an approach which can share high security information over the network using QR barcodes. QR barcodes are explained in detail to get the rigged understanding on quick response technology. The design of data security model to share the data over the network is explained. This model aims to enable secure data share over the network. This model is a layered architecture and protects the data by transforming the structure of content. So barcodes are used to put tricks over the information rather than directly using it for its noble functionality.**

Keywords-QR Barcode; Quick response technology; Data Security; Information Security; Image processing; Data Sharing Architecture

I. INTRODUCTION

The emergence of diverse networked data sources has created new opportunities for the sharing and exchange of data. The use of information has become a persistent part of our daily life. Employees use information to perform elementary job functions; managers require significant amounts of it for planning, organizing and controlling; corporations leverage it for strategic advantage.

Since the application of computers in administrative information processing is also very important, computers have become a key instrument in the development of information processing. The rapid development of information technology (IT) has helped to firmly establish the general attitude that information systems are a powerful instrument for solving problems.

Utilizing these emerging technologies, however, is not without problems. People start considering their sensitive information when it is transmitted through open networks; they began worrying about using forged information for business; and corporations worry about customer and investor confidence if they fail to protect sensitive information. Protecting sensitive information has consequently become a top priority for organizations of all sizes.

The majority of existing systems focus on performance and precision of data retrieval and information management. A number of techniques are employed to protect information; however, in many cases, these techniques are proving inadequate. For example, while several information systems use the add-ons security features to provide information confidentiality (which allow users to share information from a data media while keeping their channel private), these security measures are insufficient.

Considering barcodes as an effective media to share information, at present, the black and white two dimensional barcode technology has developed more mature. American National Standards Institute (ANSI) developed the international standards of two-dimensional bar code, QR codes [1][2]. However, with the urge of increase in information, expanding the field of bar code applications is thought to be good idea. Barcodes are used to store the high capacity information in less space and thus, stands as a good ideate.

In the upcoming section, the structure, information capacity and other important properties of QR barcodes are explained. In section 3, a model is proposed with an algorithm. Finally in the last section, future work and conclusion for the model is proposed.

II. BARCODE APPLICATIONS

An effective and innovative call from Denso introduced the world with Barcodes which are so efficient in their work, especially in specific domains like automated storage systems and data hiding [3][4].

Barcodes finds good applications in inventory control systems like stock level maintenance, control on incoming goods and raw materials, etc. Also barcodes has production control applications like status on information of authentications of goods. They are also extensively used in industries and business sectors related to automotive, food processing, electronics, insurance, pharmaceutical, and so on [5].

As if now, author does not find barcodes to be used to secure data while transferring over the network. Before knowing the proposed work using barcodes, how barcodes are invented and what is the structure behind the creation of barcodes is discussed in next section.

III. RESEARCH ON QR BARCODES

The strength of QR Barcodes can be identified as they contain the forte of PDF 417 for its high data capacity, Data Matrix for its reduce space printing and MAXI Code for its high speed reading as presented in fig. 1.

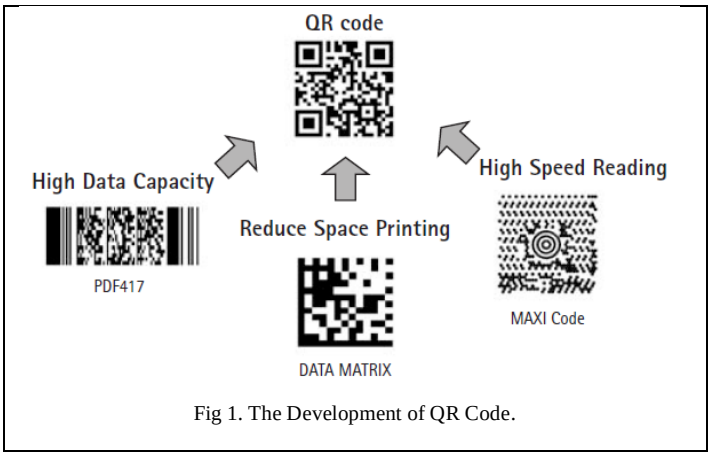


Fig 1. The Development of QR Code.

A. Generating Codes

QR codes can be generated online with various code sizes. For optimum readability we recommend a quiet zone area of 4 modules around the code. A module is the smallest pixel element of a QR Code. For most phones a smaller quiet zone of 2 modules will work, but 4 modules are required in the official documentation [8]. So we also recommend the module size to be 4 so that it captures the most information within and maximizes the capacity size. See fig. 2.

B. Encoding Process

There are three steps in the encoding process of colour bar code: encoding the data information, generating error correcting code, and generating symbol process [9]. Firstly, according to the data information encoding mode, the data is transformed into data stream. Meanwhile, the information of length of data stream is added into the head of data stream.

Then each four bits in the data stream form a code and the corresponding error-correcting code is generated through Reed-Solomon algorithm. The error-correcting code is added to the end of the data stream. Finally, all of the data codes will be transformed into symbols and the function images are added too as shown in fig. 3.

Fig. 2. QR Barcode Module.

C. Information Capacity

In the black and white two dimensional bar encoded method, the data is transformed into binary data stream [10]. Then these binary data value, 0 and 1, are presented with black and white symbols. Each module which can present 0 or 1, can express two kinds of different information. So in the black-and-white two-dimensional barcode with n symbol modules, the information capacity will be 2n.

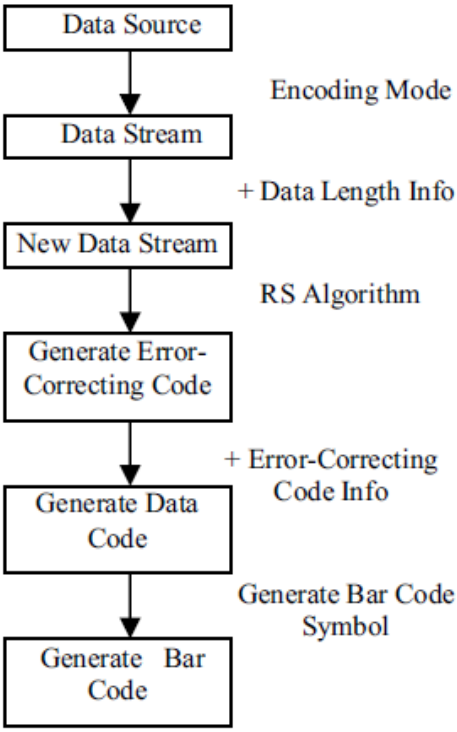


Fig. 3. The Encoding Process.

D. QR Barcode Structure

The QR barcode [6][7] is a two dimensional symbol developed by Denso Wave in 1994. The code contains information in both the x-axis and y-axis, whereas traditional barcodes contain data in one direction only. The main structure of the QR barcode is shown in fig. 4. The outer range is the quiet zone. The upper-left, upper-right, and left bottom square areas are used for position detection and pattern separators for positioning.

There are six smaller squares which are the alignment patterns. Additionally, the main area, which is colored grey, is the kernel area of data and error correction code. The QR code's size is decided by determining a symbol version based on data capacity, character type, numeric, alphanumeric, etc., and error correction level, and by setting a module size based on the target printer's or scanner's performance level. The placement of finder patterns and timing patterns can be seen in fig. 5.

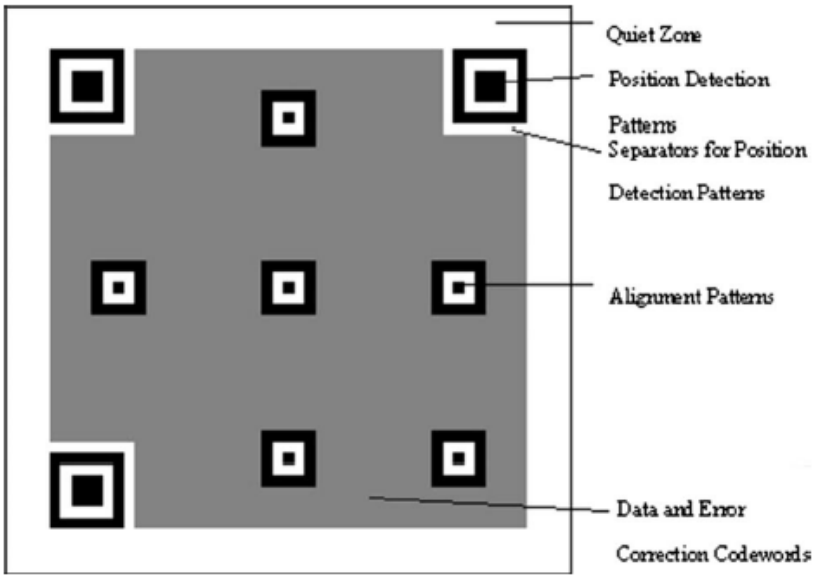


Fig. 4. The Structure of QR Barcodes.

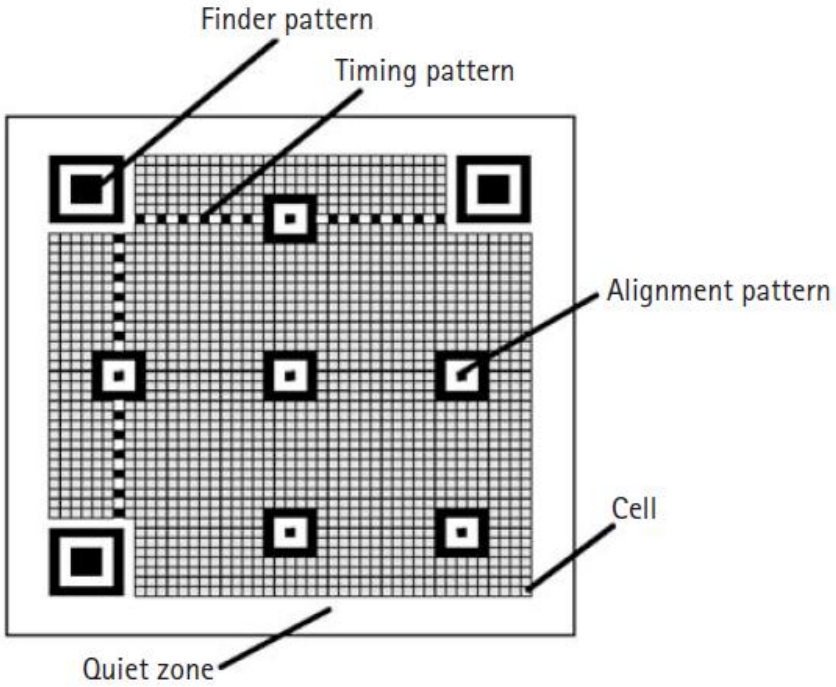


Fig. 5. The Finder and Timing Pattern of QR Barcode.

E. QR Code Structure in Detail

Fig. 6 shows an example of QR code symbol. The figure is version 1 (type 2) and the module is 21 X 21 cells, vertical 21 cells and horizontal 21 cells. This version is specified from 1 to 40, increased by 4 cells per one version up. The maximum version is 40 and the size is 177 X 177 modules. Fig. 6 is a case of the QR code version 1 modules that are arranged in a grid pattern of black and white squares. In this QR code symbol [11][12][13], there are three position detection patterns (Finder patterns) in the upper left corner, bottom left and top right corner. Then the timing pattern is placed between every one of these position detection patterns. Additionally, alignment patterns are introduced in the version 7 or higher. Then Table I shows the main specifications of the QR code. There are four modes available,

(1) number mode,

(2) alphanumeric mode,

(3) 8 bit byte mode and

(4) kanji and kana characters mode.

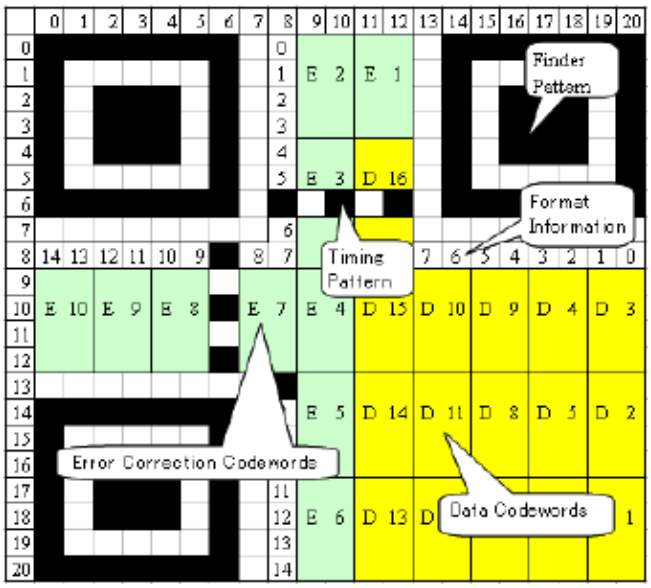


Fig. 6. Example Structure of QR Code.

It can also be combination of these modes. The RS (Reed Solomon) error correction code is used for recovering from symbol dirty or transmission error. There are four levels of error correction capability, Level L: about 7% recovery, Level M: about 15% error recovery, Level Q: about 20% error recovery and Level H: about 30% recovery.

IV. Proposed Algorithm

An algorithm is proposed here which secures the information while being shared on the network. The idea is to create a random number corresponding to the information saved in the database. The random number behaves as the key to the information in the database. Each random unique key point to the unique information placed in the database server.

Barcodes are used for the obvious reasons. They are acceptable worldwide. Various barcode readers are openly available on the internet and barcodes can be widely scanned in mobile devices.

For the random number created initially, a quick response barcode is created. QR barcode generators are widely available in the market. The preferable size of QR barcode in this algorithm is 800 length and 800 height to utilize more information space in barcodes.

The barcode created is now used to create two identical looking barcodes with minor differences in each which are not detectable by naked eyes. The identical image *one* is created by first detecting the black continuous series of pixel in barcode. Then the first black pixel from the series is formatted to white or base color. The process is repeated for all such black series of pixels in the image. This image is now named as the *False Image One*. False image one, on completion of formatting for each black series of pixel in the image, then looks identical to the original image.

Table I. Minimum QR Code Specifications.

Item	Specifications	
Error Correcting Code	RS Code	Data
	BCH Code	Format Information
		Version Information
Characters	Number	10 bit coding per 3 number digits
	Alphanumeric	11 bit coding per 2 characters
	8 bit byte	8 bit coding
	Kanji	13 bit coding per 2 characters
Version	1	21 x 21 modules
	2	25 x 25 modules
	40	177 x 177 modules
Error Correcting Level	L	about 7%
	M	about 15%
	Q	about 25%
	H	about 30%
Finder Pattern	1:1:3:1:1	3 co-centric squares
		7x7, 5x5, 3x3 modules
Alignment pattern	1:1:1:1:1	3 co-centric squares
		higher version 2
		5x5, 3x3, 1x1 modules

Now the second false image is created out of the original image. Same process as in creating false image one will be followed with a small difference. Black series of pixels are identified and then unlike *false image one*, this time the last black pixel from the series is formatted to white or base color instead of first black pixel in the series.

This is another identical looking image to the original barcode created and thus, called *False Image Two*.

As a result of steps described above, we get three images all identical to each other. But only one image which is original can be read by the *Barcode Readers*.

The false image one is then sent over the network to the destination. After a predefined scheduled time difference, false image two is also sent to the destination. The two images are scanned at the receiver's end and fed to the algorithm decoder.

Summarized Algorithm: (See fig. 7)

Step 1: Create a ***unique*** alpha numeric random number of length 1024 characters corresponding to the message to be secured.

This will allow 62^{1024} different combinations that are infinite.

Step 2: Of this random number, a QR Barcode (Quick Response) Image is generated of size 800 * 800.

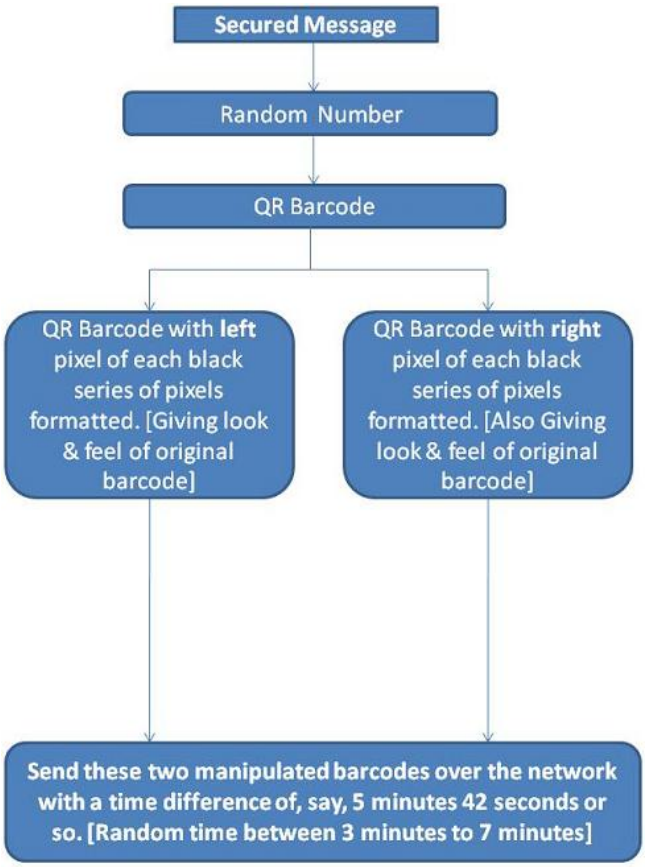


Fig. 7. Proposed Algorithm.

Step 3: Two Barcode images are created from the original barcode image each of size 800*800. One image will be manipulated by inverting one left most black pixel of each black continuous series of black pixels. Likewise, other image will be manipulated by inverting one right most black pixel of each black continuous series of black pixels. This will give the effect of three similar images.

Step 4: These two images are sent to the destination end with the time difference of 3 to 7 minutes (randomly created, on the fly) and the reverse process is followed at the destination to get the secured message.

V. Conclusion

Protection of sensitive information is a growing concern around the globe. Securing critical data in sectors like business, healthcare and military, has become the first priority of sensitive information management. Failing to protect this asset results in high costs, loss of customers and investor confidence and even *threaten national security*.

QR Codes are two dimensional barcodes that can contain any alphanumeric text. Quick Response (QR) codes are versatile. A piece of long multilingual text, a linked URL, an automated SMS message, a business card or just about any information can be embedded into the two-dimensional barcode. Barcodes are used to store the high capacity information in less space and thus, stands as a good ideate.

An algorithm is proposed in this paper which efficiently transfers the data with high security with the aim of able to protect sensitive information. This can be used for very high security data and its complexity makes it more efficient to secure the information. Thus, a competent and innovative way of utilizing the barcode technology is proposed. The system is efficient enough to handle the 1024 characters and leads to completely efficient and reliable results. Image processing makes the system's processing more complex but stands for its application purpose.

VI. Further Work in Progress

The implementation of the design of secure data exchange system is in progress. Also a real time application case study is in progress to its accomplishment. Also the capacity of the barcode can be increased by using color barcode.

References

[1] International Organization for Standardization: Information Technology International Symbology Speciation – Data Matrix [S]. ISO/IEC16022 (P), 2000.

[2] Pavlidis T, Swartz J. Fundamentals of barcode information theory[J]. IEEE Computer, 990, 23(4): 74- 86.

[3] Thota Sriram, K.V.Rao, S Biswas, Basheer Ahmed, "Application of barcode technology in automated storage and retrieval systems", BHEL.

[4] Bing Nan, Ming-Chui-Dong Y Mang, "From Codabar to ISBT 128: Imlementing Barcode Technology in Blood Banking Automation System", Proceedings of the 2005 IEEEEngineer in Medicine and Biology 27th Annual conference, Shanghai, China, 2005.

[5] Barcoding : A complete tracking system from design to despatch; Mark Marriot, Numeric Arts Ltd., (Logistics today).

[6] Wen-Yuan Chen, Jing-Wein Wang; "Nested image steganography scheme using QR-barcode technique" Optical Engineering 48_5, 057004 _May 2009_Vol. 48(5).

[7] See http://www.denso-wave.com/qrcode/qrfeature-e.html_.

[8] A Guide to Kaywa's QR code solutions, See http://mobile.kaywa.com/qr-code-data-matrix.

[9] Wicker, Bhargava. Reed - Solomon Codes and Applications [J]. IEEE Press, 1994.

[10] Zhi Liu, Herong Zheng, Huaguo Jia; "Design and Implementation of Color Two-Dimension Barcode with High Compression Ratio for Chinese Characters", International Conference on Information Engineering and Computer Science, 2009 (ICIECS 2009).

[11] Wakahara, Toshihiko; Yamamoto, Noriyasu; "Image Processing of 2-Dimensional Barcode", Conference on Network-Based Information Systems (NBiS), 2011 14th International .

[12] Japanese Industrial Standards, "Two Dimensional Symbol-QR-Code-Basic Specification" JIS X 0510, October 2004.

[13] T. J. Soo, "QR Code", Synthesis Journal, pp..59-78 2008.

[14] Xianping Wu; "Security Architecture for Sensitive Information Systems".

Authors Profile

Sona Kaushik belongs to National Capital Region, New Delhi, India. She received the B.Tech. degree in Information Technology in 2007. She is currently working as a System Engineer in a reputed IT Organisation and pursuing Masters in Technology from Birla Institute of Technology, Mesra, Ranchi, India. Her research interests are in information security, network security, security engineering and cryptography.

Graphing emotional patterns by dilation of the iris in video sequences

Rodolfo Romero Herrera
Departamento de Posgrado, Escuela Superior de Computo (ESCOM - IPN)
México D.F.

Francisco Gallegos Funes
Escuela Superior de Ingeniería Mecánica y Electrica (ESIME –IPN)
México D.F.

Saul De La O Torres
Departamento de Posgrado, Escuela Superior de Computo (ESCOM - IPN)
México D.F.

***Abstract*— For this paper, we took videos of iris of people while induced a feeling of joy or sadness, using videos to motivate the states affective. The manuscript implemented is a system of recognition affective pattern by dilating the iris, with which extracted images of the videos. The results obtained are plotted to facilitate the interpretation. A suitable treatment occurred for locating the pupil and the obtaining of the diameter of the pupil. The graphics are based on statistical time intervals.**

Within the research found that the iris diameter varies depending on your mood present in the subject of study. There is software that can detect changes in the pupil diameter, however it is also develops software, the main objective is to detect changes with respect to affective states and in this study is the main contribution. The joy and sadness were the emotional states that may differ. The system presents graphs that can be observed when analyzing the dependence between feelings and dilation present in the eye.

Keywords- Pupil dilation; joy; sadness; mathematical morphology; average interpolation.

I. INTRODUCTION

The emotion recognition is a step to truly intelligent machines, therefore is essential, for recognition the emotional pattern recognition to understand the human intelligence [1].

Based on the growing interest in using the technology to read emotions in a human (affective systems) [2] [3] [4] and the fact that they have found evidence that emotions cause changes in the iris [5], The system developed is a system to recognize and classify the patterns that occur in the iris of a person while changes your mood, focusing in two of the basic human emotions: joy and sadness. Although the reader should know that Paul Ekman bases their work on six basic emotions [6].

II. METHODOLOGY

The pupil is a hole that is images of black. In order to eliminate the illumination and brightness was used HSV color format [7][8], the brightness component is greater than 0.5, as it goes from 0 to 1; however it is worth taking a higher than average brightness. Located these points, it creates a new pixel with the same value H and S, but lower in V. MAX is the maximum value of the components (R, G, B), and MIN the minimum value, then we have that HSV by (1) (2) and (3) allow the conversion to this format [9].

$$H = \begin{cases} not\ defined, & if\ MAX = MIN \\ 60^o x \frac{G-B}{MAX-MIN} + 0^o, & if\ MAX = R \ \ y \ \ G \geq B \\ 60^o x \frac{G-B}{MAX-MIN} + 360^o, & if\ MAX = R \ \ y\ G \geq B \\ 60^o x \frac{G-B}{MAX-MIN} + 120^o, & if\ MAX = G \\ 60^o x \frac{G-B}{MAX-MIN} + 240^o, & if\ MAX = B \end{cases} \quad (1)$$

$$S = \begin{cases} 0, & if\ MAX = 0 \\ 1 - \frac{MIN}{MAX}, & in\ the\ other\ case \end{cases} \quad (2)$$

$$V = MAX \quad (3)$$

For the analysis of the eye, it is isolating the darkest elements of the image, which also are tabs and areas outside the pupil. However, these other areas are irregular, so it makes an analysis of distances of each pixel to background of the image, figure 1. All operations for isolate the iris is made in a binary image, where has a 1 if this point is considered possible pupil and 0 if not [10].

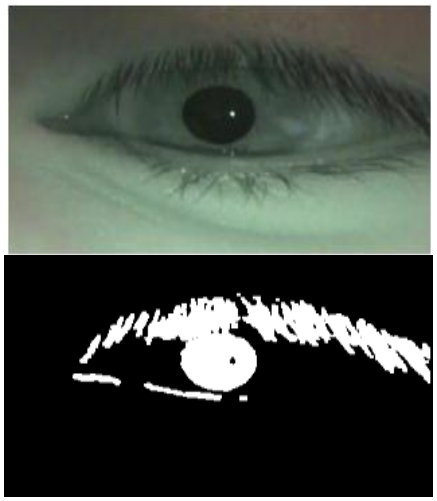


Fig. 1. Input Image and Image binarized.

The pixels with greater distance from the background are considered part of the pupil, determining the pupil area. See Figure 2.

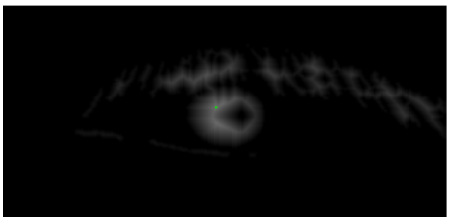


Fig. 2. Image of distances, the white represents the greatest distance from pixel to the bottom; the circle is the highest value.

By means of operations of morphology mathematical of opening are eliminated small elements (noise) (4)[5][11].

$$A \cdot B = (A \ominus B) \oplus B \tag{4}$$

Once detected the pupil, algorithms are used to fill the pupil, through the closure operation (5) [5][12].

$$A \cdot B = (A \oplus B) \ominus B \tag{5}$$

Once obtained and recorded the diameter of the pupil of each image from the video, the system allows see these results graphically as shown figures 3 to 11. In the y-axis are the mm by the pupil diameter, on the axis "X" are represents all the images from the video, it are get 20 pictures for each second of video.

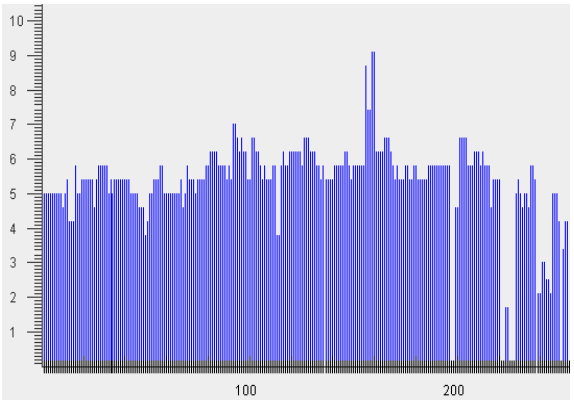


Fig. 3. Graphic bars

Fig. 4. Figure continues

In graph with bars and the continuous plot of figures 3 and 4 are shown the diameter of the pupil respectively found in each of the images [13].

In the figure 5, it was a combination of the graph with bars and graph continuous.

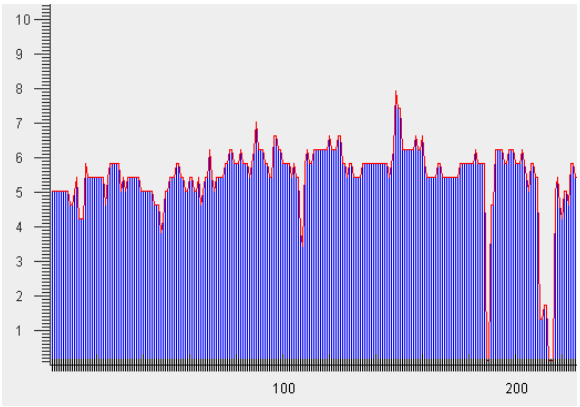


Fig. 5. Graph combination

In figure 6, red lines represent the average obtained after twenty images, that is to say, after each second video. In the graph of the Figure 7 can see how the diameter of the pupil changes in each second of video.

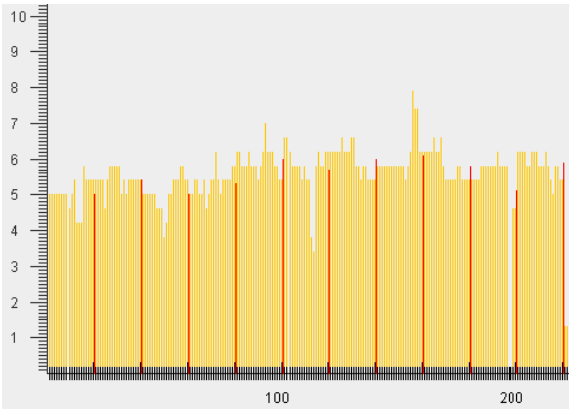


Fig. 6. Graph with average

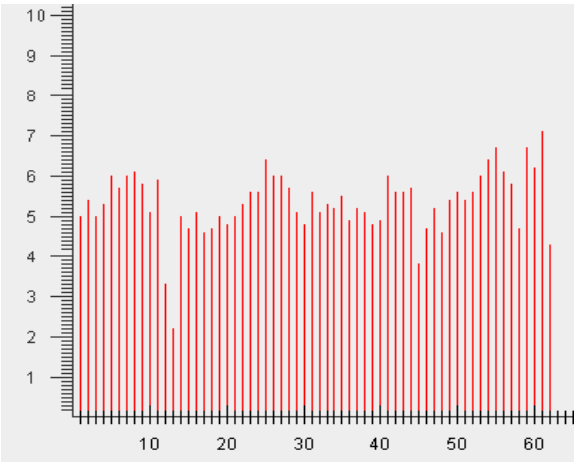


Fig. 7. Graph average.

In the graphs of figures 8 and 9, it is possible to recognize the results found after Interpol, blue bars and black line segments now represent the images that initially gave the zero.

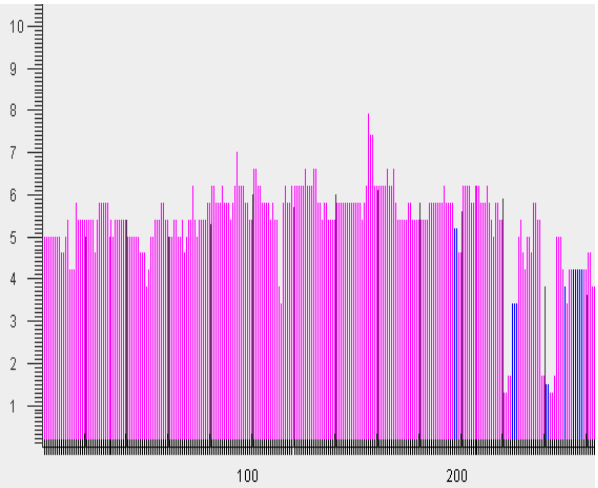


Fig. 8. Graph Interpolation bars

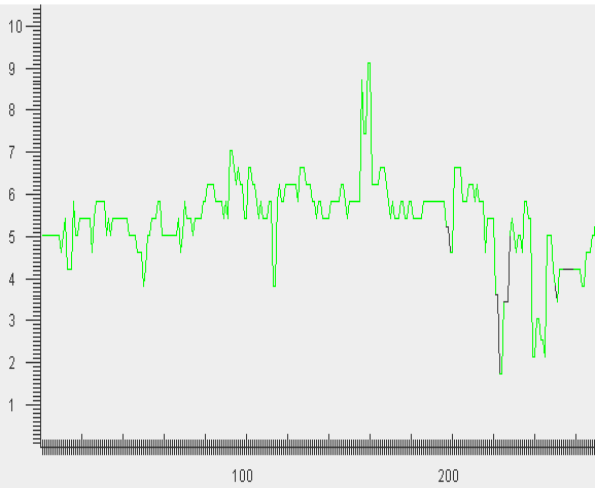


Fig. 9. Graph Continuous Interpolation

The graph in figure 10 gives the diameter obtained in each second of video, after the interpolation. In the graph of the Figure 11 is possible to show the image corresponds to the bar where is positioned and press the mouse computer.

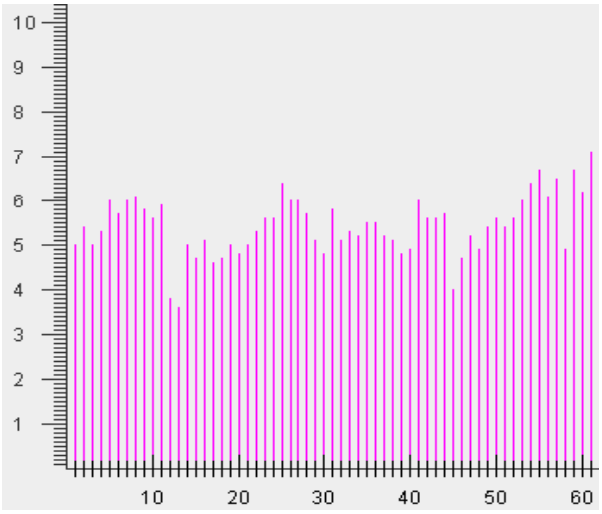


Fig. 10. Graph Average Interpolation

Fig. 11. Graphical analyzed image

III. RESULTS

The following chart shows the average diameter of the pupil for each second of video of one of the subjects of test with predominant emotion “joy”:

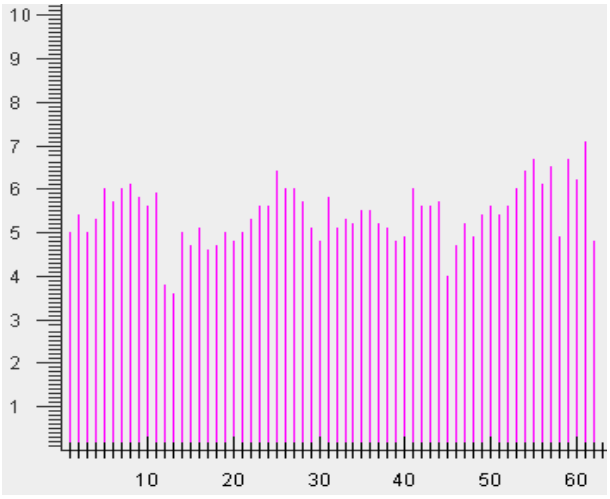


Fig. 12. Figure-1-subject-3

After analyzing the video that corresponds to the graph of figure 12, observe that when the subject makes laughter increases the amplitude.

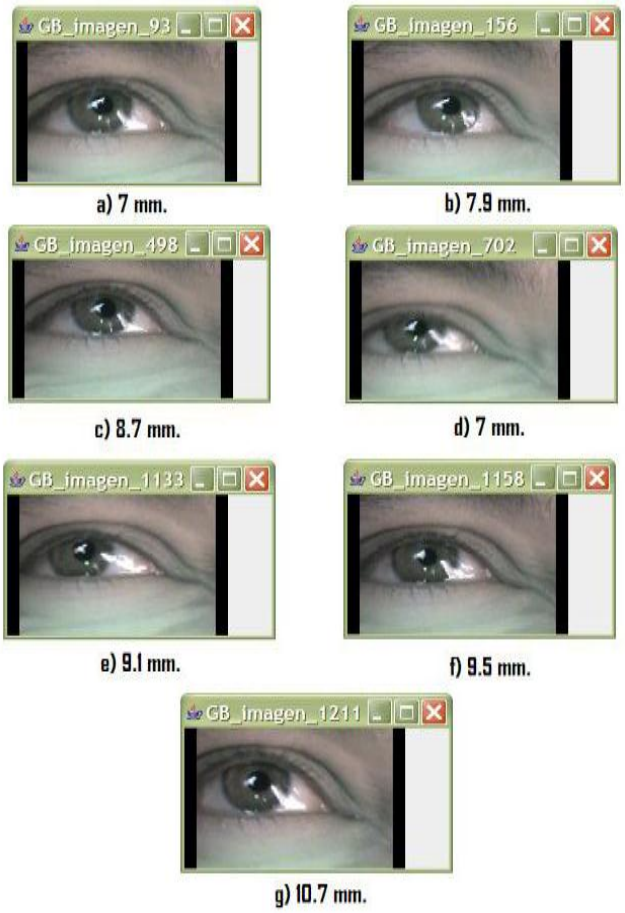


Fig. 13. Images of heights.

Figure 13 contains images that correspond to the high points of the graph bars in figure 12. The graphs show interesting results when values of zero diameter. These images are usually where the person closed the eye or the pupil; the eye is in the edge, or simply an image that does not contain an eye, figure 14 shows some examples.

Figure 14 Images zero.

The software can detect the joy and sadness through graphs, however, is detectable (but it is very difficult) other emotional states, what requires more extensive analysis to determine the mixed affective states.

The analysis allows get median, variance, standard deviation and the range where we find most of the data [14].

IV. Conclusions

Working with emotions is not easy, the fact feel recorded or observed does not allow feelings flow easily, although this is able to detect changes in the iris of individuals, some researchers have even found not only changes in the eye with the affective state, also with the degree of cognition [15] [16], but has not determined a range of dilatation in relation to affective state. The feeling where became more evident the change was the joy. It is noticeable from the moment you take the video, when someone laughs the pupil diameter increases proportionately, but not directly, on a range of 5 to 7 mm, it is possible to say that there is a dependency statistics. In some cases difficulties were encountered because many people tend to close their eyes when laugh, blink or cry, but these are changes that also the system can detect.

After a number of tests to the same person are located data that indicates which mood is reflecting this in their eyes; if stored the data the system can indicate when the subject is sad or happy, following the same procedure for each person. It is possible to consider other types of emotions like anger, stress, fear, etc.

Once the system can indicate the state of mood of a subject, we will have to see, that so congruent it is the emotion of the person with who indicates software; which is complex, because the subject may lie in their emotional state. However is clear, the mood changes the diameter of the pupil, allowing differentiate the joy of sadness.

ACKNOWLEDGMENT

Thanks to The IPN (National Polytechnic Institute) who through of COFAA (Commission for promotion to the academic support) allows us to conduct the researches.

REFERENCES

[1] Rosalind W. Picard, Alan Wexelblat, Cliffrod I. Nass ; Conference on Human Factors in Computing Systems **;** CHI '02 extended abstracts on Human factors in computing systems; Minneapolis, Minnesota, USA ; Pages: 698 - 699 Year of Publication: 2002 ISBN:1-58113- 454-1

[2] Juan D. Velásquez, Pattie Maes; International Conference on Autonomous Agents Proceedings of the first international conference on Autonomous agents; Marina del Rey, California, United States; Pages: 518 - 519 Year of Publication: 1997 ISBN:0-89791-877-0

[3] Rosalind W. Picard, Alan Wexelblat, Cliffrod I. Nass ; Conference on Human Factors in Computing Systems **;** CHI '02 extended abstracts on Human factors in computing systems; Minneapolis, Minnesota, USA ; Pages: 698 - 699 Year of Publication: 2002 ISBN:1-58113-454-1

[4] Ric Heishman, Zoran Duric, Harry Wechsler; Using Eye Region Biometrics to Reveal Affective and Cognitive States, George Mason University, 2004.

[5] Rodney Brooks; Humanoid robots; Communications of the ACM ; Volume 45 , Issue 3 ; Robots: intelligence, versatility, adaptivity ; Pages: 33 – 38,2002.

[6] Paul Ekman, Como detectar menterias, Paidos, 2009.

[7] Costantino Grana; Roberto Mezzani, Rita Cucchiara; Enhancing HSV histograms with achromatic points detection for video retrieval; Conference On Image And Video Retrieval; Proceedings of the 6th ACM international conference on Image and video retrieval ;Amsterdam, The Netherlands ;Pages: 302 - 308 ;Year of Publication: 2007 ; ISBN:978-1-59593-733-9

[8] Arturo de la Escalera; Visión por Computador Fundamentos y Métodos; Prentice Hall Madrid (2001)

[9] Aristide Chikando, Jason Kinser; Optimizing Image Segmentation Using Color Model Mixtures; 34th Applied Imagery and Pattern Recognition Workshop – Cover; Washington, DC; 2005

[10] Hugo Proenc¸a and Lu´ıs A. Alexandre; UBIRIS: A Noisy Iris Image Database; Lecture Notes in Computer Science Volume 3617,Springer;USA; 2005.

[11] M. Harlick Robert, R. Sternberg Stanlye, Xinhua Zhuang; Image Analysis Using Mathematical Morphology; IEEE Transactions on pattern Analysis and Machine Intelligence, Vol. Pam-9, No. 4, July 1997.

[12] De Mira, J, Mayer, J; Image feature extraction for application of biometric identification of iris - a morphological approach; Computer Graphics and Image Processing,SIBGRAPI 2003; Brasil, 2003.

[13] Pértega Díaz S., Pita Fernández S; Representación gráfica en el Análisis de Datos; Unidad de Epidemiología Clínica y Bioestadística; 2001

[14] L. Ipiña Santiago, I. Durand Ana; Inferencia estadística y análisis de datos; Pearson Prentice Hal; España; 2008

[15] Ric Heishman, Zoran Duric, Harry Wechsler (2004): Using Eye Region Biometrics to Reveal Affective and Cognitive States, George Mason University.

[16] Rodolfo Romero Herrera, Saúl De la O Torres, Melody Neftali Rivera Gutiérrez, (2005) ROC&C (IEEE) "Detección de cambios en el diámetro de la pupila para el reconocimiento de estados emocionales".

AUTHORS PROFILE

Master degree in computer systems and electronic communications

The impact of competitive intelligence on products and services innovation in organizations

Phathutshedzo Nemutanzhela
Faculty of Information and Communication Technology
Tshwane University of Technology (TUT)
Pretoria, South Africa

Tiko Iyamu
Faculty of Information and Communication Technology
Tshwane University of Technology (TUT)
Pretoria, South Africa

***Abstract*—This paper discusses the findings of the study that was aimed at establishing the effect of Competitive Intelligence, hereafter referred to as CI, on product and service innovations. A literature study revealed that CI and innovation are widely studied subjects, but the impact of CI on innovation was not well documented. CI has been widely acclaimed as a panacea to a lot of organizational problems. The study aimed at establishing the impact of Competitive Intelligence (CI) on products and services Innovation in organisation. A case study was conducted, using an Information and communication technology (ICT) organisation. Innovation-decision process was applied in the data analysis. At the end of the study on the impact of competitive intelligence on products and services innovation in organisations the following was achieved. It was better understood that while CI is overemphasized as revolutionary, customer focused information systems products and services still remain challenging. It was also understood that not all organisations that deploy CI produce more innovative methods. A lack of knowledge- sharing and limitations within the organisational culture were found to be important factors for the deployment of competitive intelligence products and services in the organisations. Conclusions of the study basing on the findings are presented.**

Keywords-Competitive Intelligence (CI); Diffusion of Innovation (DoI); Innovation; Product; Services.

I. INTRODUCTION

The ability to innovate is at the heart of every business' survival. When a business runs out of innovative capabilities, then it is bound to exit the fight to attain and keep customers. [1] Contends with this view that "for over 20 years, successful product innovation has been considered a key requirement for business success".

For innovation to realize its full value there is need for the market to buy the final product or service. This necessitates the innovating firm to gather as much knowledge as it can about the needs of the customer, or to create such a need in the market. This kind of creating a need is what [2] called the need pull model. In the need pull model the search for a high potential market initiates a search for the inside or outside the firm knowledge to develop an innovative product that would meet market a need. This is the usefulness of Competitive Intelligence to gather the required knowledge to create opportunity in the market.

Besides the firm creating a need, there are other needs that are created as a result of responses to the competition. [3] contends that "Competitive Intelligence units can also help a firm to understand how a rival has developed their own unique capabilities and asset caches, assess a rival's ability to imitate their strategy or assist a company to assess how to uniquely bundle resources to create value for its customers." This way a firm will reposition itself to respond with the situation in the competitive environment. This would be a direct impact of CI.

II. LITERATURE REVIEW

The study aimed to develop a framework for understanding the impact of competitive intelligence on products and services innovation in organisations. Related studies were reviewed, with consideration that this is a conference proceedings paper.

According to Cavalcanti [4] "The essence of intelligence begins with environmental scanning activities, also known as surveillance. The essence of this process is a transformation of data, information and knowledge into intelligence as a final product." However intelligence as a final product becomes useful if and when the final consumer's needs have been satisfactorily met.

Another value of Competitive Intelligence is the ability to gather consumer opinion. [4] Asserts that "Consumer opinion can offer insight into the benefits of a product or a service as well as consumption tendencies. It can also help identify wishes, dreams and future fantasies." These future fantasies are the heart of the innovative process. As soon as the organisation can gain insight of the wishes of the consumers, this will direct the innovations in an attempt to satisfy the consumer needs. The extent to which the organisations turn this information into a resource for product or service innovation is the reason of this study.

In the literature, it is customary make a distinction between (competitive) intelligence as a product and as a process (e.g., [5, 6, and 7]). In treating intelligence as a product, authors refer to the "information" or "knowledge" obtained and used for strategic purposes. The process view stresses the process by means of which this information or knowledge is obtained. Both the above definitions stress the process aspect. The first definition also highlights intelligence as a product. If competitive intelligence is seen as a product, it is usually compared with data, information and knowledge [5, 6, and 7].

CI has two basic functions namely to move competitive losses to wins and move competitive pushes (no decisions) to

wins [8]. [8] Asserts that "To achieve this kind of impact, Competitive Intelligence must address the core issues of departments that have influence on the drivers of competitive outcomes. Specifically, Competitive Intelligence needs to address mission critical issues for sales, marketing, and product development." The issue asserted here is the ability of the users of Competitive Intelligence to transform whatever information they have gathered into an innovative product or service. The best way it will address sales and marketing is providing an innovative product or service in addition to formulating a formidable strategy. Hence the purpose of the study was to establish how far CI contributes to the process of product and service innovation.

A) CI in the process of products and services innovation

According to some studies, innovation is not a single act but rather a process which begins with an idea or an invention [9, 10, 11, 12, and 13]. The classification of Innovations into different types focuses attention on the outcomes of the innovation process which have been defined as a new product or service, a new production process technology, a new structure or administrative system or a new plan or program pertaining to organisational members [9].

The power of competitive intelligence to contribute to product and service innovation lies in its ability to enable technological forecasting, as [14] aptly titled their 1998 paper "Technological forecasting techniques and Competitive Intelligence: tools for improving the Innovation process". Their argument is that "Technological forecasting can be used as a tool to help the forecasters in the process of describing, in the most precise way possible; how a future machine will be; and its method of operation and its characteristics. In other words, how, where and what should be innovated." The information that serves as the intelligence is that information that is in the right hands.

Information with a touch of expertise makes a difference in the process of Innovation.

Frates and Seena's [15] argue that an important subset of competitive intelligence, competitor intelligence is insufficient, and potentially misleading. If all a company does is track known competitors within its own industry, it is likely to suffer from marketing myopia. This implies that a firm must not only focus on its competitors, but, more importantly, the whole general business environment. In essence the goal of CI is to facilitate a whole rounded body of knowledge that an organisation can draw from in its formulation of the Innovations it will present to the customer in form of goods and services.

However [15] do not explicitly show what impact CI would have on innovating for the new customers that they refer to. While they advocate targeting new customer segments, there is a need to enter these new markets with innovative products that will capture the markets.

III. RESEARCH METHODOLOGY

The study was descriptive because it was aimed at answering the question "what". Descriptive study "tries to discover answers to the questions who, what, when, where, and sometimes, how. The researcher attempts to describe or define a subject…" [16] The aim was to establish what are the impacts of Competitive Intelligence on product and service innovation in organization are. According to Cooper and Schindler [16] "qualitative refers to the meaning, the definition or analogy or model or metaphor charactering something, while quantitative assumes the meaning and refers to the measure of it." The research had a qualitative attribute because this allowed the researcher to gather perceptions and opinion about Competitive Intelligence. The quantitative view gathered the proportions of the practitioners of Competitive Intelligence and be able to generalize their opinion. Since the study collected opinion about the impact of Competitive Intelligence in the innovation process, these opinions were recorded as qualitative data because reference would be made about them in the form of quotes. These opinions were gathered through semi-structured interviews.

There were two main, primary and secondary sources used in the data collection, namely, interviews and document review, respectively. According to Leedy and Ormrod [17] data collection may be anything, such as electronic documents that can assist the researcher to answer a research question. Interviews were conducted with identified employees at Amajita. Relevant documents about the case study were gathered from authentic sources. This includes access to the organisation's strategy.

The semi-structured interview approach was used. According to Kvale [18], the most useful interview format for conducting qualitative research is often "semi-structured" (sometimes called "moderately scheduled"). A total number of 11 interviews were conducted. This number was reached heuristically, i.e., the decision to stop adding respondents was taken when nothing new was being learnt from the interviews. The research main question was: What are the impacts of IS innovation on CI products and services in organisations? There were subsidiary question to the main question.

IV. THEORY

The Diffusion of innovations (DoI) theory was used in the analysis of the data. The notion of stages in an innovation-decision process conceptualized by [19] was employed. Therefore a model of the Innovation-decision process that covers five stages: Knowledge, persuasion, decision, implementation, and confirmation were used to gain knowledge and understanding of the impact of competitive intelligence on information systems innovation products and services in organisations. This was in order to develop a good information systems innovation framework.

The analysis was done at two different, macro and micro levels, individual and group, respectively. Each participant in the case studies was labelled as follows: in Amajita, as AJ_LA001 to AJ_LA011.

The Innovation decision process characterized as a process that occurs while individuals participate in a series of actions related to decisions [20] Knowledge occurs when individuals are aware of the Innovation and gain understanding of its functions. Persuasion is when individuals or decision-making units exhibit favourable or unfavourable behaviour toward the

Innovation. Decision indicates when the individual or unit decides to adopt or reject the Innovation. Implementation occurs when the individual or unit decides to use the Innovation. Confirmation occurs when decision makers confirm or reject their decision to adopt the Innovation [20].

V. DATA ANALYSIS

Using the Innovation-Decision Process [19], the analysis of the data is presented as follows:

A. Knowledge

Knowledge was considered as key and a very important factor in the organisation. If the organisation is to remain competitive over time, the managers have to generally see a need to know more about new products and services, technology, and about current and potential competitors. One of the reasons that were attributed to the deployment of competitive intelligence (CI) is because the market keeps changing. Also information and communication technology keep changing and rapidly. Almost on daily basis, there are a new IS innovation products and services. According to one of the interviewees AJ_L002 (p8:99-100), the constant and rapid changing requires awareness by the employees and the employers, if the organisation is to have competitive advantage. AJ_LA004 (p20:277-278) also expressed that, the market trend of IS innovation products and service always changes. The way things are done changes every day because we are in the technology world. This challenge us; managers need to research further and make sure employees are always knowledgeable about products and services we offer so that we can remain competitive in the market.

Without the knowledge of the products and services the organisation won't be able to deploy CI. According to one of the of the interviewees AJ_LA001 (p4:37-38), we send our employees to boat camps, when our partner has initiated a new product and service so that they can be educated of the challenges faced during the implementation and also get to know the benefits that the innovation bring into the organisation.

According to one of the interviewees, AJ_L003 (p32: 475-480), as an organisation, we face a challenge when employees are not aware of the products and services that are offered, because both the organisation and employees are impacted negatively. The costs of our business get affected and employee's growth is also affected. The employee further expressed that, Employees, who are not familiar with the environment they operate on, can be too reluctant and fearful to take on task. For fear of being exposed, which could lead to threat to their job, also some employees agrees or volunteer to take on task, which they lack the knowledge of, this results to poor quality of service delivery to the clients. This sometimes results in frustrations on the part of the informed employees and ultimately resignations. From the organisations point of view, these employees who lack knowledge were considered first when faced with retrenchments or dismissals.

B. Persuasion

Successful organisation consists of result-oriented persons. The capacity to achieve results consistently is tied to develop the ability to plan. Execute flawlessly, and follow-up consistently. Effective staffers require improving their personal effectiveness. It was necessary for each employee to convince themselves of the products and services. Thereafter, they could confidently implement and support the products and services on behalf of the organisation. According to one of the employees: I think as employees, we must first persuade ourselves on improving our personal effectiveness towards the deployment of CI before we can even question our managers on the measure they are putting in place to persuade us.(AJ_LA0010 p39:595-596). The employees further expressed that, even if employer put performance measure in place to persuade us, the performance will still be poor because we are not self-effective.

According to one of managers AJ_LA004 (p21:291-292), we have realized that if an employee is not ready to adopt the changes brought to them, their performance get affected as they don't become too involved in the deployment of that products and services, they even refrained themselves. For organisation to persuade employees to adopt the deployment of CI information systems products and service innovation is a challenge. This is mainly because each employee has his or her own view of how they see products and services. As such, many of the interviewees differ in their views and belief. The views and opinions were informed by individuals and groups roles in the deployment of CI in the organisation.

According to one of interviewees AJ_LA0011(p42:650-651), Our organisation get the innovation from a global partner and it then gets implemented without finding out how we feel about the innovation, all that the organisation does is to send us to training so that we can be equipped on the deployment and benefits of products and services. Therefore employees were not given room for negotiation. Whatever that gets decided by the management, the employees have to comply with it. Not having a room for negotiations was a risk to the organisation, in terms of losing employees' commitment.

AJ_LA009 (p37:558-559) asserted that, Our skills are not measured, yet our managers give incentives for performance. It doesn't motivate many employees for them to put all your effort on a work that you know everyone will be rewarded the same regardless of the effort, but if the organisation had a measure in place, I believe all employees were going to strive for the best deployment. Performance management contract was required in order to develop a process for managing individuals and teams to achieve high levels of organisational performance.

C. Decision

According to AJ_LA001 (p1:9-10), We operate on a centralised environment, and the innovation to be deployed comes from our international partner who has already gone through the whole deployment. He further expressed that our organisation has to analyse the information of IS products and service to make sure it aligns with the strategy and culture of our business, before we can decide to adopt the innovation. Therefore many of the decisions made were based on objectives and strategy of the organisation. Analysis is vital in helping decision-makers understand how phenomena in the

broader environment relate to their company's mission, objectives, and strategy.

AJ_LA001 (p4:38) mentioned that, We recognized the fact that organisations may lack the ability or means to respond or adapt to changes in the environment. This might be as a result of centralisation, which limits them innovative skills. Also the complexity and rapidity by which changes occurred had significant impact on the organisation's competitive challenges. As a result, an organisation's ability to receive or acquire information from the environment becomes an essential, yet not sufficient ingredient in the strategy formulation process. The partnership with an international company means the organisation was forced to adopt and deploy competitive intelligence in the exact way as the partners. This limits the organisation ability to decide on specific strategy. Decision-making to them, it just mostly means whether they decide to implement the product and services or not.

The operation of the organisation was centralised. Among other things, this was intended to ensure that there was uniformity of data in the organisation. AJ_LA007 (p32:480-481) stated that, The business that we operate on is a centralised operation where all information is stored and retrieved from one central place called the "cloud" so everyone get in the organisation has access to this information which then makes it easy analysis the data and make decision.

D. Implementation

The organisation preferred to use the CI as the only source of inspiration and tend to be followers rather than leaders or innovators within specific markets. According to one of the employees AJ_LA007 (p30:439-400), Our organisation doesn't come up with innovation, but rather will follow the footsteps of our partners, which limit us from being creative and think strategically. AJ_LA006 (p28:417-418) stated that, It looks easy and simple to implement products and services which are already done somewhere, but as the organisation we face challenges of aligning the innovation with our own culture and strategy.

The implementation of CI products and services is influenced by many factors. The factors include: the negative attitudes of managers, the organisations corporate culture not being conducive for CI. In some cases, we the managers tend to think that we know it all. As a result, we fail to listen to advices from our colleagues, and believe whatever we decide on has to be implemented exactly that way, the time we realize it's wrong. The damage has been done and all the money used for implementation has been wasted (AJ_LA003 p11:79-80).

In the organisation, the implementation of CI was a challenge. This challenge was attributed to lack of strengths and weakness in context of skills, lack of resources such as infrastructure, and limited cost. According an interviewee, Like another competitive intelligence implementation, our organisation sometimes face challenges of not having enough resources, sometimes we try to cut implementation cost and end up missing the important steps of implementation (AJ_LA001 p3: 22-23).

AJ_LA002 (p10:92-93) Emphasizes on implementation being the most important stage on this organisation. The employees view implementation as the most important stage on competitive intelligence deployment. From strategy to the final deployment of CI, the organisation adopt a package from USA which already in use. Which then leaves them with the ability to only implement the product and service in the South African environment after adoption? The value that competitive intelligence add during the implementation process its cutting cost and time spent on it. Amajita achieve this by using CI tools that helps them better their services and product delivery.

E. Confirmation

The confirmation of CI products and services in the organisation was done on case by case. Competitive Intelligence implementation provides multiple tangible and intangible benefits to the organisation. However, the realisation of these benefits depends on how CI was implemented. According to one of employees: the acceptance of our products and services depend on how well we implement them, some of them get accepted well, and we see this through customers that organisation acquire after deployment (AJ_LA005 p23:314-315).

The most critical step towards the successful of a CI solution is an understanding of its users. One of the interviewees AJ_L009 (p36:546-547), questioned that if feasibility study was done accurately and all customer needs were understood before implementation, the organisation had faced less challenges to whether the customer will accept this new innovation.

After the implementation of CI on IS products and services innovation is done, the whole team awaits to see how the innovation will be accepted or confirmed as a success. The organisation also go through the own confirmation stage. We test our innovation and make sure everything is working according to the project plan (AJ_LA005 p15:359-360). Even though we have assurance that the innovation works, we still need a confirmation from our local customer to see if it works also for them (AJ_LA007 (p32:477- 478).

VI. FINDINGS

From the analysis of the data, eight factors were found to be critical and influence the deployment of competitive intelligence products and services in the organisation. The factors include cultural values, knowledge sharing, environmental scanning, education and awareness, output, performance contract, centralisation and negotiation. These factors are discussed as follows:

a) Cultural Values

The arena of cultural values differs from one environment to the other. Within cultural settings, people share values about what is acceptable or not. In this sense, it is very rare for two organisations which are operating on two different continents to share the same cultural values. Within such context as Amajita find itself, this limits the creative skills of those who their organisation is not dominant. Therefore their culture and values are compromised to suit the deployment of CI products and services from another environment. But if this was not the case, the employees were going to contribute and explore in the meantime practicing and following their organisation cultural values. When cultural and values are compromised sometimes

customer lack confidence on their organisational ability to deliver.

b) Knowledge sharing

Some of information, skills, knowledge and expertise are being shared in the organisation aimed to improve and sustain CI products and services. However, there are certain factors which the organisation hasn't addressed. This factors impacts on employee's willingness to share information. Such factors are performance contract, organisation perform on a centralized environment, and fear of losing their jobs if they share knowledge and skills they have. Their environment limits and forces the employees to restrain from sharing knowledge, as they lack measure to use for persuasion. Employees who have information that is worthy to the organisation, and refrain from sharing knowledge because of fear that they might lose their jobs. They also fear that their junior could use it to compete with them, and fear of losing power that they hold. Not being in position to innovate, was also another reason why employees don't share information that can help change the scope of the business of the organisation. They wait to be told what to do. Though, organisation has realized that knowledge constitutes a valuable intangible asset for creating and sustaining competitive advantages. Knowledge sharing activities were generally supported by knowledge management systems as adopted by the organisation. However, technology constitutes only one of the many factors that affect the sharing of knowledge in organisations. Others includedorganisational culture, trust, and incentives. The sharing of knowledge constitutes a major challenge in the field of knowledge management because some employees tend to resist sharing their knowledge with the rest of the organisation. And maybe if they have measure put in place this can persuade them to share information. However, trust goes with security, when employees don't feel secure at their organisation, they won't have trust. Personal interest will always supersede the interest of the organisation. As that continue to happens, no one will be will to share any of expertise that they have.

c) Environmental scanning

The organisation depends on their international partner for innovation, limiting them to exploit of CI tools. The organisation uses forms of workshops to check whether their clients could adopt the innovation which was presented to them by the international partner. Therefore the relationship they have built with their external customer plays a bigger role during environmental scanning. When there is new innovation, they act as quickly as possible. They should be able to realize the gap immediately. So that they can either adopt or reject the innovation. During that time they can utilize their customer as the tool to acquire information, instead of hosting events that can cost money and consume time. They have to use them as tool for finding of information about the environment and to check if the innovation can fit in. This enables the organisation to understand the external forces of change so that they may develop effective responses which secure or improve their position in the future. Environmental scanning had influence on the implementation of competitive intelligence products and services. As a result of the influence, some competitive intelligence products and services suffer rejection from potential customer. This is because the organisation didn't take time to scan the environment, however just implement.

d) Education and Awareness

Education and Awareness plays a significant role in encouraging and enhancing people's participation in the activities which were aimed and considered essential for achieving sustainable competitive intelligence products and services. The organisation seems to have less emphasis on educating their staff about the information systems products and services, particularly, the implementation process of the deployment of CI products and services. It seems as they just assume that all employees are aware of the deployment process. Employees who are not aware of the organisation they operate to, they are more of an expense, than an asset to the company. Because they won't even be able to transfer or acquire skills from each other as they don't have an idea of what's happening around their organisation. However, if the organisation can spend time educating and making employees aware of the CI deployment, they can cut cost and reduce time spent doing feasibility studies on the process of adopting and implementing innovation from outside partner.

e) Output

In order to improve the effectiveness and efficiency of CI products and services, the organisation must consist of result oriented persons. The capacity to achieve results consistently was tied to development of the ability to plan, execute flawlessly, and follow-up consistently. Unfortunately, many employees in the lacked the ability to plan as they only wait to be told what to do, instruction from the international partner. This inability to innovate affects their personal effectiveness. They can only execute innovation that has already been executed by the international partner. So they were not needed to be creative and plan positively. In majority of the cases, answers were provided to them. However, an organisation is required to improve their employee's personal effectiveness. Each personnel need to have a good understanding and motivating strategies competitive intelligence products and services in the organisation. Because an employee who fails do discipline themselves affects the performance of the company.

f) Performance

Organisation need to have contract that can sets out the terms of performance between the employees and the employer. Because the organisation doesn't have performance contract in place, employees seem to lack motivation, and this had impact on CI products and services. They don't trust colleagues with certain information, skills, knowledge and expertise they have. This also led to some employees not sharing information with some of their colleagues. Some of them even resigned because they feel that they are not secure. The performance contract builds relationship between employees and the employer. Because the rights and obligations of all parties involved in the performance were covered.

g) Centralisation

The activities of the organisation regarding planning decision-making they were concentrated within a particular group. The centralisation approach had impact on innovation of

CI products and services in the organisation. When everything is centralised in the organisation, the organisation lacks skills to innovate. This means employee get to be told what to do. Employees don't have a view on the CI deployment but only to implement. The organisation force employees to reserve their input, because everything runs on a centralized space. Information and innovation comes from centre point, in most cases, from the international partner.

h) Negotiation

Employees don't have bargaining power. What employer decides on it goes. The nature of the business (centralised organisation) limit both employer and employees an opportunity to negotiate or come up with different views that could benefit the organisation. Even though the organisation is dominated by the global market, the culture and values they use are of the organisation. This put the employees on an uncomfortable situation where they are deploying a global innovation, yet the culture and values are not the same and still can't negotiate. If employees were given a chance to negotiate they could have had better way to deploy CI products and services, to suit the culture and values of the South African environment.

The above findings were interpreted and presented in the following section.

Interpretation of the Findings

The findings from the analysis are interpreted to be the factors which influence the innovation of competitive intelligence (CI) within the information systems environment. This is depicted in Figure 1 below. The discussions that follow explain the manifestation from these factors.

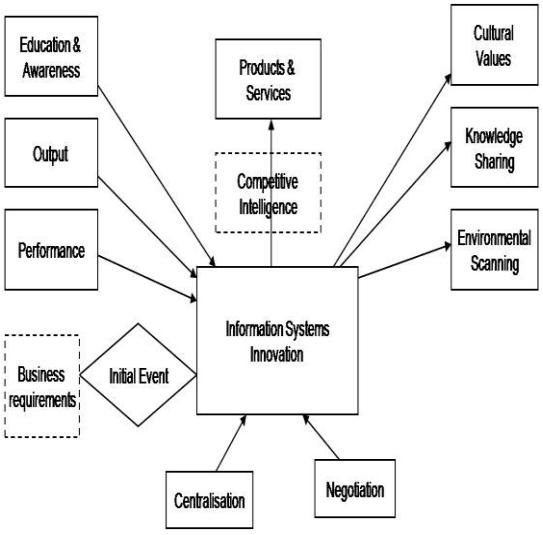


Figure 1. Factors Influencing Information Systems Innovation

a) *Improvement* – is a change in the thought process for doing something, and useful application of new inventions or discoveries. The organisation doesn't innovate but rather implement innovation, the environment and nature of the business that they operate in doesn't give them an opportunity to come up with innovation, or a discovery of any information systems products or services. When employees are not aware of the innovative process (products and services deployment) they become reluctant and lack innovative and creative thinking. The employees' acts, behaviour and performance manifest to the organisational performance. This has impact on skills and knowledge transfer.

b) *Dominant* – the international partner to Amajita was dominant in the innovation process. They are the one who comes up with innovation. It's their strategy, plans and execution that led to the deployment of the competitive intelligence, making them to be dominant party over Amajita. While the technology industry dominant technology and organisation cycles are faster than in many other organisations, the point is still that dominant organisations have to evolve or they will fail. Therefore, Global innovator should provide the full array of competencies needed by successful CI practitioners. Therefore Amajita leaders and managers must understand the significance of these sources in South African market, rather than agree to the naive concept that they get from their international partners as the market cannot be exactly the same.

c) *Distribution* - An organisation or set of organisations (go-betweens) involved in the process of making a product or service available for use or consumption by a consumer or business user. By sharing knowledge about products and services, employees gain more than they lose. Sharing knowledge is a synergistic process – you get more out than you put in.

d) *Collaboration* - is a process where two or more people or organisations work together to realise shared goals, (this is more than the intersection of common goals seen in co-operative ventures, but a deep, collective, determination to reach an identical objective). If you try to work alone – you are likely to fail – you need not only the input from other people but their support and buy-in. Being open with them; sharing with them, helps you achieve your objectives.

e) *Reliance* – employees might trust transferring skills to others, and know that the other employees and the organisation at large would benefit from it. They should also trust that the information won't be used against them tomorrow. Which means organisation has to assure them of security and that they job won't be at stake if they are willing to share information and expertise they have.

f) *Involvement* – is the state of a user in relation to a situation after that user has participated in the situation. It is, approximately, an internal analogue to aconflict of interest. Even with the low level of knowledge sharing that goes on today – if you do not make your knowledge productive than someone else with that same knowledge will. You can almost guarantee that whatever bright idea you have someone else somewhere in the organisation will be thinking along the same lines.

VII. CONCLUSION

It was established that the role of Competitive Intelligence on product and service innovation was to inform strategic management, reflect customer needs, and inform rivals about the competitors and help firms locate themselves on the competitive scale.

It was established Competitive Intelligence is overemphasized as revolutionary yet customers still remain unsatisfied by the services and products because most of Competitive Intelligence remains in talk, but not in execution. This was reflected by the numbers of respondents who said that Competitive Intelligence was not useful because what was collected was not used.

The findings and the analysis of the study indicate that further research relating to competitive intelligence can be conducted. Some of the suggestions are:

Organizational Culture – it would be in the interest of academic organisations to investigate and gain a better understanding of how organisational culture impacts on Competitive Intelligence products and services in organisations.

Underpinning theories – it would be a significant contribution to apply different theories such as Organisational information processing theory (OIPT) and Technology-Organisation-Environment Framework (TOT framework) in a similar study. Organisation need to be knowledgeable about the Competitive intelligence products and services they deploy so that they can make better decisions to innovate. Organisational information processing theory is applied to Organisations that need quality information to cope with environmental uncertainty and improve their decision making [21].

REFERENCES

[1] R. L. Chapman, C. E. O'mara , S. Ronchi S., A., M. Corso,. Continuous product innovation: a comparison of key elements across different contingency sets. Measuring Business Excellence. 2001 Pp 16-23. Available from www.emerald-library.com/ft [Accessed 10/08/2010]

[2] A. R. Burgelman, L. R. SAYLES. Transforming invention into innovation: the conceptualization stage in Burgelman A. R., christensen C. M., wheelwright S. C. 2004. Strategic management of technology and innovation 4Ed. McGraw Hill-Irwin Boston.

[3] S. Hughes. Competitive Intelligence as Competitive Advantage: The Theoretical Link between Competitive Intelligence, Strategy and Firm Performance. Journal of Competitive Intelligence and Management . Vol 3. No 3. Available at www.scip.org [Accessed 20/08/2010]

[4] P. E. Cavalcanti. The Relationship between Business Intelligence and Business Success Journal of Competitive Intelligence and Management .Vol 3. No 1 available at www.scip.org [Accessed 20/08/2010]

[5] L.M. Fuld. The New Competitor Intelligence. Chichester. UK: Wiley, 1995.

[6] L. Kahaner. Competitive Intelligence. New York: Touchstone, 1997.

[7] E. A. Philips, and D. Vriens. Business Intelligence. Deventer: Kluwer, 1999.

[8] R. B. Jeppsen. An introduction to impact-based Competitive Intelligence. 2005 www.scip.org [Accessed 16/ 05/10]

[9] F. Damanpour. Organizational innovation: a meta-analysis of effects of determinants and moderators. Acad. Mgmt. J.1991. pp.34:555.90.

[10] D. Dougherty, and E. Bowman. The effects of organizational downsizing on product innovation. California Management Review,1995. pp.37: 28-44.

[11] F. Drucker and F. Peter. The discipline of innovation, Harvard Business Review,80,(8). 2002. pp.77-83.

[12] Tidd, Joseph, Pavitt, Keith (contributor), Tidd, Joe, R. Bessant. ManagingInnovation : Integrating Technological, Market and Organizational Change, 2nd edition, John Wiley & Son Ltd, UK. 2001.

[13] A. H. Van de ven, Polley, E. Douglas, Garud, Raghu, Venkataraman, Sankaran.The Innovation Journey, Oxford University Press, USA.2000.

[14] Â. D. Lemos, and A. C. Porto. Technological forecasting techniques and Competitive Intelligence: tools for improving the Innovation process Industrial Management & Data Systems. 1998. pp7:330–337.

[15] J. Frates and S. Seena. Using Business Intelligence to Discover New Market Opportunities, Journal of Competitive Intelligence and Management, vol. 3, no. 2, 2005. available at www.scip.org [Accessed 20/03/2010]

[16] D. R. Cooper and P. S. Schindler. Business research methods 8th ed. McGraw Hill, Boston. 2003.

[17] P. D. Leedy and J. E. Ormrod. Practical research: planning and design. 8th ed.: Pearson. MAREE K., 2007. First steps in research. Pretoria: Van Schaik.2009.

[18] S. Kvale. Interviews: An introduction to qualitative research London: Sage.1996.

[19] E. M. Rogers. Diffusion of Innovations. 4th ed. New York: Free Press. 1995.

[20] E. M. Rogers. Diffusion of innovations (5th ed.). New York: Free Press. 2003.

[21] G. Premkumar, K. Ramamurthy and C. S. Saunders. Information processing view oforganisations: An exploratory examination of fit in the context of interorganisational relationships. Journal of Management Information Systems, 22(1), 2005.pp.257-294).

[22] Phathutshedzo Nemutanzhela and Tiko Iyamu; "A Framework for Enhancing the Information Systems Innovation: Using Competitive Intelligence".

AUTHORS PROFILE

Phathutshedzo Nemutanzhela has a Masters degree (MTech) Business Information Systems from Tshwane University of Technology. She has a Baccalaureus Technologies (BTech): Information Technology (Informatics). Her principle research interest is Competitive Intelligence, Information Systems and innovation. Theoretically, she focuses on Diffusion of Innovation (DoI) Theory.

Tiko Iyamu is a Professor of Information Systems at the Tshwane University of Technology, Pretoria. His research interests include Mobile Computing, Enterprise Architecture and Information Technology Strategy. Theoretically, he focuses on Actor Network Theory (ANT) and Structuration Theory (ST). Iyamu is author of numerous peer-reviewed journal and conference proceedings articles.

Arabic Sign Language (ArSL) Recognition System Using HMM

Aliaa A. A.Youssif , Amal Elsayed Aboutabl, Heba Hamdy Ali
Department of Computer Science, Faculty of Computers and Information
Helwan University
Cairo, Egypt

***Abstract*—Hand gestures enabling deaf people to communication during their daily lives rather than by speaking. A sign language is a language which, instead of using sound, uses visually transmitted gesture signs which simultaneously combine hand shapes, orientation and movement of the hands, arms, lip-patterns, body movements and facial expressions to express the speaker's thoughts. Recognizing and documenting Arabic sign language has only been paid attention to recently. There have been few attempts to develop recognition systems to allow deaf people to interact with the rest of society. This paper introduces an automatic Arabic sign language (ArSL) recognition system based on the Hidden Markov Models (HMMs). A large set of samples has been used to recognize 20 isolated words from the Standard Arabic sign language. The proposed system is signer-independent. Experiments are conducted using real ArSL videos taken for deaf people in different clothes and with different skin colors. Our system achieves an overall recognition rate reaching up to 82.22%.**

Keywords-Hand Gesture; Hand Tracking; Arabic Sign Language (ArSL); HMM; Hand Features; Hand Contours.

I. INTRODUCTION

Singing has always been part of human communications [1]. For millennia, deaf people have created and used signs among themselves. These signs were the only form of communication available for many deaf people.Within the variety of cultures of deaf people all over the world, signing evolved to form complete and sophisticated languages. These languages have been learned and elaborated by succeeding generations of deaf children.

Normally, there is no problem when two deaf persons communicate using their common sign language. The problem arises when a deaf person wants to communicate with a non-deaf person. Usually both will be dissatisfaction in a very short time.

In this section we focus our discussion of the efforts made by researchers on sign language gesture recognition in general and on Arabic sign language (ArSL) in particular. Sign language recognition systems can be further classified into signer-dependent and signer-independent. Also one may classify. Sign language recognition systems are either glove-based which relies on electromechanical devices for data collection, or none glove-based if free hands are used. The learning and recognition methods used in previous studies to recognize sign language include neural networks and hidden Markov models (HMMs).

Cyber gloves have been widely used in most of previous Works on sign language recognition including [1, 2, 3]. Kudos [4] reported a system using power gloves to recognize a set of 95 isolated Australian sign languages with 80% accuracy. Grobel and Assan [5] used HMM to recognize isolated signs with 91.3% accuracy out of a 262-sign vocabulary. They extracted 2D features from video recordings of signers wearing colored gloves. Colored gloves were used in [6] where HMM was employed to recognize 52 signs of German sign language with a single color video camera as input. In a similar work [7] an accuracy of 80.8% was reached in the corpus of 12 different signs and 10 subunits using the K-means clustering algorithm to get the subunits for continuous sign language recognition. Liang and Ouhyoung [8] employed the time-varying parameter threshold of hand posture to determine end-points in a stream of gesture input for continuous Taiwanese sign language recgnation with the average recognition rate of 80.4% over 250 signs. In their system HMM was employed, and data gloves were taken as input devices.

The use of cyber gloves or other means of input devices conflicts with recognizing gestures in a natural context and is very difficult to run in real time. Therefore, recently researchers presented several sign recognition systems based on computer vision techniques [9, 10, 11]. Starner et al [12] used a view-based approach for continuous American continuous sign language recognition. They used a single camera to extract two-dimensional features as the input of HMM. The accuracy of 92% or 98% was obtained when the camera was mounted on the desk or in a user's cap in recognizing the sentences with 40 different signs. However, the user must wear two colored gloves (a yellow glove for the right hand and an orange glove for the left) and sits in a chair before the camera. Vogler and Metaxas [14] used computer vision methods and interchangeably AT a flock of birds for 3D data extraction of 53 signs for American Sign Language. They, respectively, built context-dependent HMM and modeled transient movement to alleviate the effects of movement epenthesis. Experiments over 64 phonemes extracted from 53 signs showed that modeling the movement epenthesis has better performance than context-dependent HMM. The system provided overall accuracy of 95.83% is reported.

Furthermore, most of the above systems are signer dependent systems. A more convenient and efficient system is the one that allows deaf users to perform gestures naturally with no prior knowledge about the user. To the best of our knowledge there are a few published works on signer-independent systems. Vamplew and Adams [13] proposed a signer-independent system to recognize a set of 52 signs. The system used a modular architecture consisting of multiple feature-recognition neural networks and the nearest neighbor classifier to recognize isolated signs. They reported a recognition rate of 85% in the test set. Again the signer must wear cyber gloves while performing gestures. Another attempt is made by Fang et al [14] in which they used the SOFM/HMM model to recognize signer-independent CSL over the 4368 samples from 7 signers with 208 isolated signs.

ArSLs are still in their development stages. A glove-based and singer-dependant Arabic sign recognition system has been developed by M. Mohandes and S. I. Quadri, M. Deriche [15]. They used a data set of 15 samples for each of the 300 signs which were carried out by a signer wearing a pair of colored gloves (orange and yellow) achieving recognition accuracy about 88.73%. Jarrah and Halawani [3] developed a system for ArSL alphabet recognition using a collection of Adaptive Neuro-Fuzzy Inference Systems, a form of supervised learning. They used images of bare hands instead of colored gloves to permit the user to interact with the system conveniently. The used feature set comprised lengths of vectors that were selected to span the fingertips' region and training was accomplished by the use of a hybrid learning algorithm achieving recognition accuracy of 93.55%. Likewise, Assaleh and Rousan [1] extended the work in [3] by using Polynomial classifiers extracted superior results on the same dataset. Their work required the participants to wear gloves with colored tips while performing the gestures to simplify the image segmentation stage. They extracted features including the relative position and orientation of the fingertips with respect to the wrist and to each other. The resulting system achieved 93.41% recognition accuracy.

More recently, recognition of video-based isolated Arabic sign language gestures is reported by in [16] and [17]. In [16], the dataset is based on 23 gestures performed by 3 signers. The data collection phase did not impose any restrictions on clothing or background. Forward or bi-directional prediction error of the input video sign was accumulated and threshold into a single image. The still image is then transformed into the frequency domain. The feature vector that represents the gesture is then based on the frequency coefficients. Simple classification techniques such as KNN, linear and Bayesian were used. This work was extended in [17] where a block-based motion estimation technique was used to find motion vectors between successive images. Such vectors are then rearranged into intensity images and transformed into the frequency domain.

This paper is organized as follows. Section II describes the Arabic sign language database used in the work. Section III describes the hand features. Hand tracking and recognition phases of the proposed system are elaborated in Section IV. Section V presents the modeling of the Arabic Sign Langue using Hidden Markov Models. The experiments and results are discussed in Section VI. Finally, conclusion and future work are presented in Section VIII.

II. Arabic Sign Language (ArSL) Database

Because there has been no serious attention to Arabic sign language recognition, there are no common databases available for researchers in this field. Therefore, we had to build our own database with reasonable size. As depicted in Fig. 1, in the video capturing stage, a single digital camera was used to acquire the gestures from signers in a video format. At this stage, the video is saved in the AVI format in order to be analyzed in later stages. When it comes to recognition, the video is streamed directly to the recognition engine.

Figure 1. Sample videos of our Arabic sign language (ArSL) database.

The database in this work is collected in collaboration with ASDAA' Association for Serving the Hearing Impaired (ASDAA) [18]. The videos are captured from the deaf community who volunteered to perform the signs to generate samples for our study. The database consists of a 20-word lexicon given in Table 1.

No restrictions are imposed on the signer or word length. The words pertain to common situations in which handicapped people might find themselves in. The database itself consists of 45 repetitions of each of the 20 words performed by different signers, 20 of which are used for training and 18 for testing. No restriction is imposed on clothing, background, age or sex of the signer. Moreover, signers are gloves-free and with different signers with different skin colors. It was totally free hands. The deaf sign word is captured using a digital video camera. The frame rate was set to 25 frames per second with a spatial resolution of 640x480.

III. Hand Features

The image features, together with information about their relative orientation, position and scale, are used for defining understated but discriminating view-based object model [19].

We represent the hand by a model consisting of (i) the palm as a coarse scale blob, (ii) the five fingers as ridges at finer

scales and (iii) finger tips as even finer scale blobs as in Fig. 2. We then define different states of the hand model, depending on the number of open fingers.

TABLE I. RRECOGNITION RATES OF DIFFERENT HMM MODELS WITHDIFFERENT FEATURE VECTOR LENGTHS

Word Number	Arabic Meaning	English Meaning
1	اسكندرية	Alexandria
2	يشرب	Drink
3	يأكل	Eat
4	انا	Me
5	انت	You
6	لغه انجليزية	English
7	جمعية	Association
8	صم وبكم	Deaf
9	لغه عربية	Arabic
10	لغه فرنسية	French
11	كتاب	Book
12	سعيد	Happy
13	ثرثار	Talkative
14	حائر	Confused
15	غيرمستريح	Uncomfortable
16	مدرسة	School
17	مسئول	Responsible
18	مصر	Egypt
19	منزل	Home
20	ينام	Sleep

To model translations, rotations and scaling transformations of the hand, a feature vector is defined to describe different hand features including the global position y), size and orientation and its discrete state.

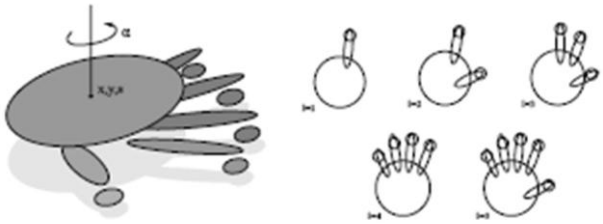


Figure 2. Feature-based hand models in different states. The circles and ellipses correspond to blob and ridge features. When aligning models to images, the features are translated, rotated and scaled according to the feature vector

IV. HAND TRACKING AND RECOGNITION PHASES OF PROPOSED SYSTEM

In this paper, a system for recognizing Arabic sign language gestures is presented. There are three main phases for hand detection and tracking; skin detection, edge detection and hand fingertips tracking.

A. Skin Detection

Each video contains a collection of frames representing a gesture. At first, each video is pre-processed by applying a video segmentation technique that captures frames with a frame rate of 25Hz. Then, the RGB captured frames are converted into HSV image because it is more related to human color perception [20]. These color spaces separates three components: the hue (H), the saturation (S) and the brightness (V). Essentially, HSV-type color spaces are deformations of the RGB color cube .They can be mapped from the RGB space via a nonlinear transformation. One of advantages of these color spaces in skin detection is that they allow users to specify the boundary of the skin color class in terms of the hue and saturation. As V gives the brightness information, they are often dropped to reduce illumination dependency of skin color.

Given an image, each pixel in the image is classified as a skin or non - skin using color information. The histogram is normalized and if the height of the bin corresponding to H and S values of a pixel exceeds a threshold called skin threshold (obtained empirically), this pixel is considered a skin pixel. Otherwise, the pixel is considered a non-skin pixel. A general image and its skin detected image can be seen such that white pixels represent the hand gesture and black pixels represent the background or any object behind the skin as shown in Fig. 3 (a). Finally, smoothing is applied to each frame using a MEDIAN filter to remove noise and shadow.

B. Canny Edge Detection

The Canny algorithm uses an optimal edge detector based on a set of criteria which include finding the most edges by minimizing the error rate, marke edges as closely as possible to the actual edges to maximize localization and marke edges only once when a single edge exists for minimal response [21].

C. Hand Contours and Fingertips Tracking

Hand tracking is the process of locating a moving hand (or both hands) over time using a camera. For each frame extract, the contours of all the detected skin regions in binary image using connected component analysis are detected. Tests are performed to detect whether the input contour is convex or not. The contour must be simple, i.e. without self-intersections. The signer's head is considered to be the biggest detected region and the moving hand as the second biggest region. Features considered include the position of the head, coordinates of the center of the hand region and direction angle of the hand region. Other features that represent the shape of the hand are also considered and are extracted from changes of image intensities called image motion:

$$I(x, y, t + T) = I\big(x - \varepsilon(x, y, t, T), y(n(x, y, t, T)\big) \qquad (1)$$

Thus, the next frame recorded at time $t + 1$ can be obtained by moving every point in the current frame, recorded at time t, by suitable amount. The amount of motion $\delta = (\varepsilon, n)$ is called displacement of the point at $X = (x, y)$.

The displacement vector is a function of the image position X , and variations in it are often noticeable even within the small tracking window. We try to find interesting points with big eigenvalues in an image to be added to the feature vector.

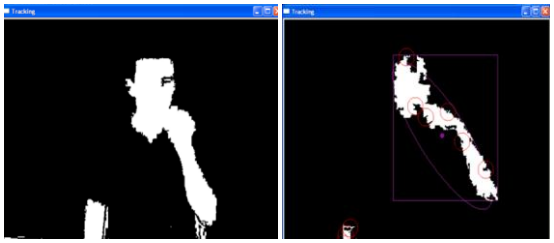
These interesting points (corners) are characterized by a large variety in all directions of the vector . By analyzing the eigenvalues of the image pixels, this characterization can be expressed in the following way: we should have two "large" eigenvalues for an interesting point. Based on the magnitudes of the eigenvalues, the following inferences can be made based on this argument:

If λ1≈0 and λ2 ≈0 (the two "interesting points" eigenvalues for an inter, st point) then this pixel (x,y) has no features of interest. In this case we reject the corners with the minimal eigenvalue less than quality Level

If λ1 and λ2 have large positive values, then a corner is found. The Shi-Tomasi [22, 23] corner detector directly computes min (λ1, λ2) because under certain assumptions, the corners are more stable for tracking.

Finally, it ensures that all the corners found are distanced enough on from another by considering the corners (the strongest corners are considered first) and checking that the distance between the newly considered feature and the features considered earlier is larger than the minimum distance. So, the function removes the features than are too close to the stronger feature.

An example of interesting points found that represent a motion is shown in Fig. 3(b).In this work, the implementation of Hand tracking method is carried out using OpenCV (Open Source Computer Vision); a library of programming functions for real time computer vision. [24].



(a) (b)

Figure 5. the result of computing blob features and ridge features from an image of a hand. (a) Result image after skin detection (b) circles and ellipses corresponding to the significant blob and features extracted from an image of a hand; it describes how the selected image features capture the essential structure of a hand.

V. MODELLING OFARSL USING HMM

HMMs (Hidden Markov Models) have been prominently and successfully used in sign languages. HMM is a probabilistic model representing a given process with a set of states (not directly observed) and transition probabilities between the states. Such a model has been used in a number of applications including the recognition of the Sign Language recognition [25, 26].

Let each sign be represented by a sequence of gestures or *observations O*, defined as:

$$O = o1, o2, \ldots\ldots\ldots, o\tau \qquad (2)$$

Where o_t is the feature vector observed at time *t*. The sign recognition problem can then be regarded as that of computing:

$$\arg\max_i \{P(w_i|O)\} \qquad (3)$$

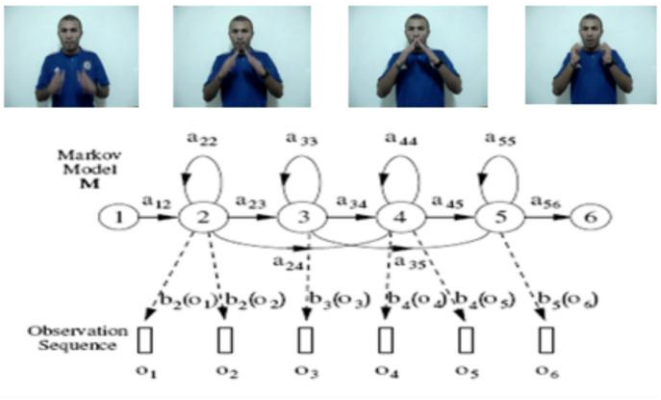


Figure 3. State HMM model for gesture "اسكندريهAlexandria"

Where w_i is the *i*'th vocabulary word. This probability is not computable directly but using Bayes' Rule [27]:

$$\Pi(\omega_i|O) = (\Pi(O|\omega_i)\Pi(\omega_i))/\Pi(O) \quad (4)$$

Thus, for a given set of prior probabilities P (wi), the most likely sign depends only on the likelihood P (O|wi). Given the dimensionality of the observation sequence O, the direct estimation of the joint conditional probability P (o1, o2 …|wi) from examples of sign is not possible. However, if a parametric model of word production such as a Markov model is.

As shown in Fig 4, each gesture for a sign is modeled as a single HMM with N observations per gesture (o1, o2, … ot). In HMM based sign recognition, it is assumed that the sequence of observed feature vectors corresponding to each gesture is generated by a Markov model as shown in Fig 4. A Markov model is a finite state machine which changes state once every time unit and each time t that a state j is entered, a feature vector ot is generated from the probability density bj(ot).

Furthermore, the transition from state i to state j is also probabilistic and is governed by the discrete probability aij . Fig 4 shows an example of this process where the six state model moves through the state sequence X = 1; 2; 2; 3; 4; 4; 5; 6 in order to generate the sequence o1 to o6. It is to be noted that the entry and exit states of a HMM are non-emitting in the Hidden Markov Model Toolkit which is used in this work [28]. This facilitates the construction of composite models as explained in more detail later.

The joint probability that O is generated by the model M moving through the state sequence X is calculated simply as the product of the transition probabilities and the output probabilities. So is for the state sequence X in Fig 4.

A. Training Phase

Training in the context of our work means learning or generating an HMM given a sequence of observations. For each training sequence XT11 , ...,XTn1 , ...,XTN1 of a gesture of class k with N sequences, the image features are prepared and then extracted .

These extracted images are used as feature vectors in the Viterbi [15] training to train a hidden Markov model λk for each gesture as shown in Fig. 5.

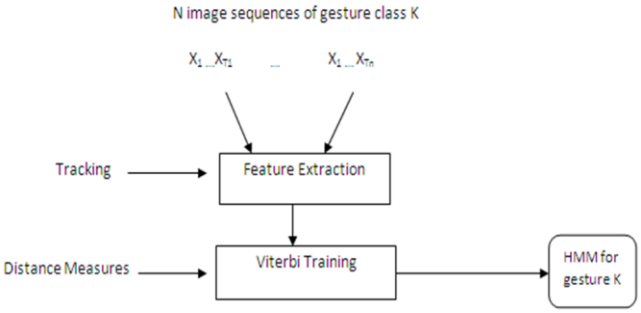


Figure 5. Traning phase

B. Recognition Phase

This phase involves finding the probability of an observed sequence given an HMM and finding the sequence of hidden states that most probably generated an observed sequence. The feature extraction of the test sequences is identical to the training process. Then for each test pattern the hidden Markov model which best describes the current observation sequence is searched as shown in Fig. 6.

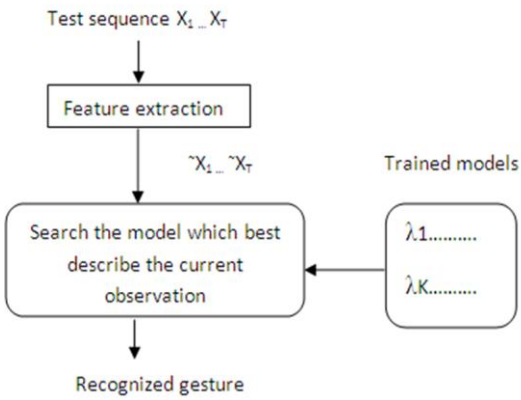


Figure 6. Recognition Phase

The implementation of the HMM for our ArSL system has been carried out using the HTK toolkit [29]. HTK is a portable toolkit for building and manipulating hidden Markov models. HTK is primarily used for speech recognition research although it has been used for numerous other research areas including speech synthesis, character recognition and DNA sequencing. HTK is in use at hundreds of sites worldwide.

VI. Experiments and Results

The training method, described earlier, has been implemented by creating one HMM model per class (gesture), resulting in a total of 20 models.

Several experiments have been conducted to evaluate our ArSL recognition system. All experiments are performed on the same training data collected in prior. Depending on the same database, the system attempts to recognize all samples for every word where the total number of samples considered here is 360. We use 6 HMM models which have different number of states and different number of Gaussian mixtures per state. For each experiment, the 6 models are used for different feature vector lengths (5, 8 and 9) achieving recognition rates of 78.61%, 82.22% and 80.27 % respectively as shown in table 2.

TABLE II. Recognation Rates of Different HMM Models with number of states and mixture

HMM model	No of features		
	5 elements	*8 elements*	*9 elements*
[3states/10mixture]	47.78%	58.89%	56.67%
[6states/2mixture]	41.94%	45.83%	55.56%
[4states/2mixture]	34.44%	30%	30%
[6states/6mixture]	61.94%	66.11%	68.61%
[6states/4 mixture]	56.94%	56.94%	43.05%
[6states/10mixture]	75.55%	71.94%	74.16%
Overall Recognition Rate	78.61%	82.22%	80.27%

The set of experiments shown in table 2 has been conducted for each of the 20 Arabic signs in our database. The best result (recognition rate) obtained for each sign along with the associated best model is shown in table 3.

It is noticeable that some signs result in particularly low recognition rates. The gesture "eat", for example, has a recognition rate of 55.56% and is mostly classified as "drink". This is due to the fact that the location, movement and orientation of the dominant hand are very similar in both gestures. Therefore, the observation (feature) vectors, o1, o2,……., ot produced from the feature hand tracking phase are most likely very close to each other. Thus, the system will get confused between these two signs and provide relatively higher error rate for these particular gestures. A similar situation occurs with the sign "Me" which has a recognition rate of 66.67% is mostly classified as {You}.

Our ArSL proposed recognition system is based on 8 features per frame which is considered better than the previously published results in the field of ArSL while Jarrah and Halawani [3] use of 30 elements as a length of feature vector per video frame. M. Rousan and K. Assaleh [15] use feature vector of 50 elements.

We compare our work with that done in [28]. In spite of the fact that they use different feature extraction methods, setup, and database, both systems are based on the same classifier (HMM). As shown in Table 4, our proposed system (referred to as ArSL-Using HMM in Table4) system performs much better than the DCT coefficient-based system (referred to as the ArSL – DCT coefficient-based system in Table 4).

ArSL– DCT coefficient-based system uses 50 elements length of feature vector and Recognition rate 90.6%.It is expected that the increase in a feature vector size be accompanied by a corresponding increase in recognition rates. This is due to the fact that each DCT coefficient is uncorrelated with other coefficients and hence no redundant information is presented in increasing coefficients.

TABLE III. RECOGNITION RATES WITH DIFFERENT MODELS FOR EVERY WORD WITH BEST MODEL

,Arabic Meaning	English Meaning	Best Result	HMM [Number of States / Number of Gaussian mixtures]
اسكندرية	Alexandria	77.78%	[6states/10 mixture]
يشرب	Drink	88.89%	[3states/10 mixture]
ياكل	Eat	55.56%	[3states/10 mixture]
انا	Me	83.33%	[6states/10 mixture]
انت	You	66.67%	[6states/10 mixture]
لغه انجليزية	English	88.89%	[6states/10 mixture]
جمعية	Association	66.67%	[6states/6 mixture]
صم وبكم	Deaf	88.89%	[6states/10 mixture]
لغه عربية	Arabic	88.89%	[6states/6 mixture]
لغه فرنسية	French	94.44%	[6states/4 mixture]
كتاب	Book	55.56%	[6states/10 mixture]
سعيد	Happy	100%	[6states/6 mixture]
ثرثار	Talkative	72.22%	[6states/10 mixture]
حائر	Confused	100%	[6states/10 mixture]
غيرمستريح	Uncomfortable	88.89%	[6states/6 mixture]
مدرسة	School	77.78%	[6states/10 mixture]
مسئول	Responsible	100%	[6states/10 mixture]
مصر	Egypt	77.78%	[6states/6 mixture]
منزل	Home	72.22%	[6states/6 mixture]
ينام	Sleep	100%	[3states/10 mixture]

TABLE IV. COMPARISON WITH SIMILAR SIGNER-INDEPENDENT, HMM-BASED SYSTEMS.

	Instruments used	Feature vector length	Recognition Rate
ArSL–DCT coefficient-based system [28]	None: free hands	50 elements of DCT coefficients per frame	90.6%
ArSL-Using HMM	None: free hands considering the head position	8 features per frame	82.22%

VII. CONCLUSION AND FUTURE WORK

We have demonstrated that our Arabic sign language recognition system is effective considering the nature of the videos used and the number of features considered. Our system is signer-independent. The database that we built consists of videos taken for deaf people using their normal life Arabic sign language. The signers are gloves-free, with varying clothes and skin colors. Importantly, only 8 features have been considered which is less than the number of features used previously by other researchers. The overall recognition rate is 82.22%, which is reasonably high considering the number of features used.

In the future, we aim to achieve higher recognition rates with a larger data set. A psycholinguistic study on the structure of Arabic sign language might be needed to choose the appropriate HMM model (the perfect number of states) for each gesture. We will explore and test our training models to build a continuous sentence recognition system using a sub-gesture word based recognition system. Such a system will help the deaf community to interact and integrate with the rest of the society.

REFERENCES

[1] K. Assaleh and M. Al-Rousan, "Recognition of Arabic Sign Language Alphabet Using Polynomial Classifiers" EURASIP Journal on Applied Signal Processing, 2005 (13): 2136-2146,2005.

[2] W. Gao, J.Y. Ma, J.Q. Wu, C.L. Wang, Sign language recognition based on HMM/ ANN/DP, International Journal of Pattern Recognition Artificial Intelligent 14 (5) (2000) 587–602.

[3] O. Al-Jarrah, A. Halawani, Recognition of gestures in Arabic sign language using neuro-fuzzy systems, Artificial Intelligence 2 (133) 117–138, 2001.

[4] M.W. Kadous, Machine recognition of Auslan signs using PowerGloves: towards large-lexicon recognition of sign language, in: Proceedings of the Workshop on the Integration of Gestures in Language and Speech, pp. 165–174, 1996.

[5] K. Grobel, M. Assan, Isolated sign language recognition using hidden Markov models, in: Proceedings of the International Conference on System, Man and Cybernetics, pp. 162–167, 1997.

[6] H. Hienz, B. Bauer, K.F. Krais, HMM-based continuous sign language recognition using stochastic grammar, in: Proceedings ofGW'99, LNAI 1739, pp. 185–196, 1999.

[7] B. Bauer, K.F. Kraiss, Towards an automatic sign language recognition system using subunits, in: Proceedings of the International Gesture Workshop, pp.64–75, 2001.

[8] R.H. Liang, M. Ouhyoung, A real-time continuous gesture recognition system for sign language, in: Proceedings of the Third International Conference on Automatic Face and Gesture Recognition, pp. 558–565, 1998.

[9] V.I. Pavlovic, R. Sharma, T.S. Huang, Visual interpretation of hand gestures for human–computer interaction: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) 677–695, 1997.

[10] J.J. Triesch, C. Malsburg, A system for person-independent hand posture recognition against complex backgrounds, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (12) 1449–1453, 2001.

[11] P. Vamplew, A. Adams, Recognition of sign language gestures using neural networks,Australian Journal of Intelligent Information Processing Systems 5 (2) 94–102, 1998.

[12] T. Starner, J. Weaver, A. Pentland, Real-time American sign language recognition using desk and wearable computer-based video, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (12) 1371–1375, 1998.

[13] Y. Wu, T.S. Huang, Vision-based gesture recognition: a review, in: Proceedings of the International Gesture Workshop, pp. 103–115, 1999.

[14] G.L. Fang, W. Gao, J.Y. Ma, Signer-independent sign language recognition based on SOFM/HMM, in: Proceedings of the IEEE ICCV Workshop Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems, pp. 90–95, 2001.

[15] M. Mohandes, S. I. Quadri, M. Deriche "Arabic Sign Language Recognition an Image–Based Approach" 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07) 0-7695-2847-3/07 $20.00 © 2007.

[16] T. Shanableh, K. Assaleh and M. Al-Rousan, "Spatio- Temporal Feature-Extraction Techniques for Isolated Gesture Recognition in Arabic Sign Language" IEEE Trans. On Systems, Man and Cybernetics Part B, 37 (3): 641-650, 2007.

[17] T. Shanableh and K. Assaleh, "Telescopic Vector Composition and polar accumulated motion residuals for feature extraction in Arabic Sign Language recognition" EURASIP Journal on Image and Video Processing, vol. 2007, Article ID 87929, 10 pages, 2007.

[18] ASDAA' Association For Serving The Hearing Impaired (ASDAA)

[19] Lars Bretzner1, 2 Ivan Laptev1, Tony Lindeberg "Hand Gesture Recognition using Multi-Scale Colour Features, Hierarchical Models and Particle Filtering" Proceedings of the Fifth IEEE International

Conference on Automatic Face and Gesture Recognition (FGR.02) 0-7695-1602-5/02 $17.00 IEEE © 2002

[20] A. Albiol, L. Torres, and E. J. Delp. "Optimum color spaces for skin detection." In proceedings of the 2001 international conference on image processing, volume 1, vol. 1, pp. 122-124 ,2001.

[21] Ravikiran J, Kavi Mahesh, Suhas Mahishi, Dheeraj R, Sudheender S, Nitin V Pujari "Finger Detection for Sign Language Recognition" Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, Hong Kong, 2009.

[22] J. Shi and C. Tomasi, "Good Features to Track,". 9th IEEE Conference on Computer Vision and Pattern Recognition. Springer, June 1994.

[23] C. Tomasi and T. Kanade, "Detection and Tracking of Point Features". Pattern Recognition 37: 165–168, 2004.

[24] Open Source Computer website. [Online]. Available Vision http://opencv.willowgarage.com/

[25] Prof. Dr.-Ing. H. Ney, "Appearance-Based Gesture Recognition" Diplomarbeit im Fach Informatik Rheinisch-Westf¨alische Technische Hochschule Aachen Lehrstuhl f¨ur Informatik VI, 2005

[26] Khaled Assaleh, Tamer Shanableh, Mustafa Fanaswala, Harish Bajaj, and Farnaz Amin, "Vision-based system for Continuous Arabic Sign Language Recognition in user dependent mode" Proceeding of the 5th International Symposium on Mechatronics and its Applications (ISMA08), Amman, Jordan, May 27-29, 2008.

[27] Forney GD, "The Viterbi algorithm". Proceedings of the IEEE 61 (3): 268–278. doi:10.1109/PROC.1973.9030, 1973.

[28] M. AL-Rousan, K. Assaleh , A. Tala'a. "Video-based signer-independent Arabic sign language recognition using hidden Markov models" Applied Soft Computing 9 990–999, 2009.

[29] The HTK website. [Online]. Available http://htk.eng.cam.ac.uk

AUTHORS PROFILE

Aliaa A. A.Youssif, professor of computer science. in Faculty of Computers and Information, Helwan University. Cairo, Egypt. She received her B.Sc and MSc. degree in telecommunications and electronics engineering from Helwan University. Dr. A. Youssif received the PhD degree in computer science from Helwan University in 2000. She was a visiting professor at George Washington University (Washington DC, USA) in 2005. She was also a visiting professor at Cardiff University in UK (2008). Her fields of interest include pattern recognition, AI researches, and medical imaging. She published more than 40 papers in different fields.

Amal Elsayed Aboutabl is currently an Assistant Professor at the Computer Science Department, Faculty of Computers and Information, Helwan University, Cairo, Egypt. She received her B.Sc. in Computer Science from the American University in Cairo and both of her M.Sc. and Ph.D. in Computer Science from Cairo University. She worked for IBM and ICL in Egypt for seven years. She was also a Fulbright Scholar at the Department of Computer Science, University of Virginia, USA. Her current research interests include parallel computing, performance evaluation and image processing.

Heba Hamdy Ali received her BSc. degree in computer science from Helwan University. She is currently a master's degree student under supervision of prof. Aliaa A. A.Youssif and Dr Amal Elsayed Aboutabl. Her areas of interests include image processing, pattern recognition, HMM and artificial intelligence.

Modularity Index Metrics for Java-Based Open Source Software Projects

Andi Wahju Rahardjo Emanuel
Informatics Bachelor Program,
Faculty of Information Technology,
Maranatha Christian University,
Bandung, Indonesia

Retantyo Wardoyo, Jazi Eko Istiyanto,
Khabib Mustofa
Dept. of Computer Science and Electronics,
Universitas Gadjah Mada,
Yogyakarta, Indonesia

***Abstract* — Open Source Software (OSS) Projects are gaining popularity these days, and they become alternatives in building software system. Despite many failures in these projects, there are some success stories with one of the identified success factors is modularity. This paper presents the first quantitative software metrics to measure modularity level of Java-based OSS Projects called Modularity Index. This software metrics is formulated by analyzing modularity traits such as size, complexity, cohesion, and coupling of 59 Java-based OSS Projects from sourceforge.net using SONAR tool. These OSS Projects are selected since they have been downloaded more than 100K times and believed to have the required modularity trait to be successful. The software metrics related to modularity in class, package and system level of these projects are extracted and analyzed. The similarities found are then analyzed to determine the class quality, package quality, and then combined with system architecture measure to formulate the Modularity Index. The case study of measuring Modularity Index during the evolution of JFreeChart project has shown that this software metrics is able to identify strengths and potential problems of the project.**

Keywords-Open source software projects; modularity; Java; sourceforge; software metrics; system architecture.

I. INTRODUCTION

Open Source Software (OSS) Projects are gaining popularity these days. They were once only considered as an experimental way of academics and researchers to share the programming experiences, now they become the mainstream software development methodology comparable to those of commercial and proprietary software projects. This movement was initially started by Richard Stallman [33] and Eric Raymond [31]. Some success stories of OSS Projects include Linux Operating System, Apache Web Server, Mozilla Web Browser, LibreOffice, etc. The success of these projects is attributed to many key success factors such as the fact that the developer is the actual user [10], and sound and modular architecture [20][17][11], the existence of communities that support the system development [9], etc. From all these success factors, modularity of the software system is one of the important factors to be examined further in this paper.

Even though there are some proofs of the success of OSS Projects, some facts that many more similar projects are unsuccessful or failed also unavoidable exist [16]. There are some characteristics of OSS Projects that have been identified contributing to such unfruitful result such as no formal means i.e. no project planning [4], poor coding styles of project initiators [13] and poor architectural design [12]. We believe that some new approaches with respect to modularity to counter such problems in OSS Projects are needed. Until now, modularity has been identified as a key success factor of OSS projects, but how to apply modularity, especially from early phase of the project is not yet understood.

This paper presents the formulation of Modularity Index which is the first quantitative software metrics to measure the modularity level in OSS Projects.

This paper is organized as follows: section 2 describes the recent studies in OSS Projects, modularity in OSS Projects and Software Metrics. Section 3 describes the data source of OSS Projects for analysis. Section 4 shows the step by step Modularity Index formulation starts from class level, package level, and system level. The case study of 33 out of 52 versions of JFreeChart projects is shown in section 5. Finally, section 6 describes the conclusion of the paper and future studies of the research.

II. RECENT STUDIES

A. OSS Projects

Many web portals have been developed as an incubator for OSS Project's developers to develop and host their projects. These portals are equipped with many development tools and statistics to assist the project initiator or administrator in improving their projects and other interested contributors to join the projects. Some of the popular portals are Sourceforge.net, freshmeat.net, launchpad.net, and Google Code.

The OSS Projects themselves have several distinct characteristics not found in commercial / proprietary software development [10][26], which are:

- The source code of the application is freely available for everybody to download, improve and modify [31].
- People who contribute to the development of the OSS projects are usually forming a group called communities. The recruitment process if this groups are completely voluntary [9]. This communities is an example of true merit-based system of hierarchy [11]

- The development methods of the projects are lacking of formal methodology found in commercially developed software applications [4]. The two most important activities are fixing bugs and adding features [3].

There are already many studies relating to OSS Projects that are classified into three main categories. The first category is the study of large and successful OSS Projects to find their success characteristics such as Debian [32], FreeBSD [12], Apache [27], Open BSD [22], and many more. The second category is the study to find similarities in several OSS Projects such as Apache dan Mozilla [26], 15 OSS Projects [35], and 2 OSS Projects [6]. The last category is the study on the process aspects in OSS Projects such as Requirement Engineering [30], code fault [22], Design Pattern [18], reliability model [37], phase of development [34], and work practice in OSS projects [10].

Current studies about OSS Projects mostly focus on the already successful and large projects that have already established hierarchy and system, while most of the failed and unsuccessful OSS Projects are usually small or medium sized projects [16]. The application of these hierarchy and system in already established projects into small to medium sized projects may not be suitable. In our initial research, we have conducted analysis on more than 130K OSS Projects to find their success factors [15].

B. *Modularity in OSS Projects*

Modularization involves breaking up of an software system into smaller, more independent elements known as module [23]. Booch has defined modularity as the property of a system whose modules are cohesive and loosely-coupled [24]. Fenton stated that modularity is the internal quality attribute of the software system [24]. It is also known that modularity is directly related to software architecture, since modularity is separation of a software system in independent and collaborative modules that can be organized in software architecture [29]. Modular software has several advantages such as maintainability, manageability, and comprehensibility [28]. Moreover, modularity has been identified as one of the key success factors in OSS Projects [20][17][11].

There are five attributes closely related to modularity in software system which are coupling / dependency, complexity, cohesion, and information hiding [21][7]. To have an ideal modular software system, the system should have the following attributes:

- Small size in each module (package) and many modules in the system [36]: each module / package should only responsible for simple feature, and the more complex features should be composed of many of these simple features. The possible software metrics to measure size are NCLOC (non-commenting lines of code), Lines, or Statements.
- Low coupling / dependency [5]: minimization or standardization of coupling / dependency e.g. through standard format i.e. published APIs [2], elimination of semantic dependencies, etc.
- Low complexity: hierarchy of modules that prefers flatter than taller dependency [28][2].
- High cohesion [21]: high integrity of the internal structure of software modules which is usually stated as either high cohesion or low cohesion.
- Open for extension and close to modification [5]: capability of the existing module to be extended to create a more complex module. And avoid changing already debugged code. The creation of new modules should be encourage using available extension and not modifying the already tested module.

Even though modularity is already identified as the key success factor in OSS Projects, the justification for it in large and succesful OSS Projects is purely qualitative. The software metrics attributing to the modularity properties are all separated and not yet integrated into a single measure. This paper will present a single measure called Modularity Index that quantitatively determines the modularity level of OSS Projects.

C. *Software Metrics*

Software metrics are defined as certain values which are expressed in some units attributed to software application [25]. The software metrics are useful in indicate the current state of the software and enable to compare and predict the current achievement of software applications [25]. There are several known software metrics based on its categories [25]:

- Size-related software metrics: NCLOC, Memory footprint, Number of classes / headers, Number of methods, Number of attributes, Size of compiled code, etc.
- Quality-related software metrics: Cyclomatic complexity, Number of states, Number of bugs in LOC, Coupling metrics, Inheritance metrics, etc.
- Process-related software metrics: failed builds, defect per hour, requirement changes, programming time, number of patches after release, etc.

There are currently more than 200 metrics with many different purposes [25], and one of the study by the authors are the statistical analysis of software metrics affecting modularity in OSS Projects [14].

III. DATA SOURCE OF OSS PROJECTS

The data source of the OSS Projects for the experiment is from the sourceforge.net portal since it is the largest OSS Portal.

A. *Assumptions and Considerations*

There are several consideration and assumption in selecting which OSS Projects to be analyzed, which are:

- The OSS projects are build using Java programming language, and a single package in the project resembles a “module” in modular software system. The addition of package in the software is intended as the addition of new feature in the system.

- The project's size is limited to small-to-medium-sized OSS Projects. The limitation of the size (NCLOC) of OSS Projects being evaluated are 170K. The concept of modularity is a lot easier to comprehend in object-oriented programming language (i.e. C++, Java, etc.) compared to procedural programming (i.e. C, Fortran, etc.), since the concept of module, coupling, cohesion, etc. are more straightforward. Java-based OSS Projects are selected since they are among the mostly popular object oriented programming for developing Open Source Software [16].
- The Projects should already be downloaded more than 100,000 times. This high number of downloads may indicate the "success" of the projects, which in turn may imply modularity traits that already identified as the success factor of OSS Project [20][17][11].
- The source code of the OSS Project is syntax error-free and compile-able. The SONAR tool requires that the source code should be compiled first using compile tool such as maven, or ant. Many of the OSS Projects provides separate binary and source code and it is difficult to create binary directly from the source code due to several reasons such as compile error, build tool configuration error, syntax error, etc.

B. Selected OSS Projects

Table 1. shows the list of OSS Projects as a subject for this research. The initial OSS Projects to be evaluated are 209 projects, but only 59 which are suitable to be evaluated using SONAR due to the assumptions and considerations stated in section III.A. There are total 1885 modules / packages being measured from these 59 OSS Projects.

TABLE I. LIST OF 59 SELECTED OSS PROJECTS

No	Project Name	No	Project Name
1	FreeMind	31	Jin client for chess servers
2	jEdit	32	SAX: Simple API for XML
3	TV-Browser - A free EPG	33	jKiwi
4	JFreeChart	34	Data Crow
5	JasperReports - Java Reporting	35	Wicket
6	OpenProj - Project Management	36	Cewolf - Chart TagLib Project
7	HyperSQL Database Engine	37	DrawSWF
8	yura.net	38	c3p0:JDBC DataSources / Resource Pools
9	JabRef	39	JavaGroups
10	FreeCol	40	OmegaT - multiplatform CAT tool
11	jTDS - SQL Server and Sybase JDBC driver	41	FreeGuide TV Guide
12	Torrent Episode Downloader	42	Eteria IRC Client
13	FindBugs	43	MeD's Movie Manager
14	PMD	44	subsonic
15	JGraph Diagram Component	45	kXML
16	ANts P2P	46	Jaxe
17	Paros	47	The JUMP Pilot Project
18	ProGuard Java Optimizer and Obfuscator	48	Aglet Software Development Kit
19	TripleA	49	Antenna
20	JSch	50	CBViewer
21	Jajuk	51	Sunflow Rendering System
22	FreeTTS	52	Thingamablog
23	A Java library for reading/writing Excel	53	BORG Calendar
24	checkstyle	54	Directory Synchronize Pro (DirSync Pro)
25	httpunit	55	Java Treeview
26	JMSN	56	Java Network Browser
27	PDFBox	57	Red Piranha
28	JBidwatcher	58	Cobertura
29	JTidy	59	Jake2
30	Jena	-	-

C. Steps

In order to be able to analyze these OSS Projects, there are some steps being performed, which are:

- Compile the source code using available build tool (Ant or Maven2).
- Execute maven2 script to start analyze the OSS Projects using SONAR tool.
- Creating custom portal to perform the required analysis.
- Analyze and find the correlation and similarities of all the projects such as using scatter graph, least square fit, histogram, etc.

IV. MODULARITY INDEX FORMULATION

The formulation of modularity index will start from the class level, then move up to the package level, and finally concluded in the system level.

A. Class Level Modularity

There are four software metrics that determine the level of modularity in class level, which are:

- Size Metrics which consists of: NCLOC, Lines, and Statements. NCLOC is the number of non-commenting lines of code. The selection of NCLOC will also represent the other size metrics [14].
- Cohesion: LCOM4 or Lack of Cohesion Method version 4, this version is better for object oriented programming such as Java as proposed by Hitz and Montazeri [19] which is the improvement of LCOM1 Chidamber and Kemerer [8].
- Complexity: McCabe's Cyclomatic Complexity [22] is one example of complexity metrics that widely used.

Our previous paper have shown that the size metrics and complexity metrics are highly related so this metrics may be ignored [14].

- Functions: the number of functions / methods in the class. This may indicates the complexity

1) NCLOC: Figure 1 shows the histogram of the class vs. NCLOC of the all OSS Projects being evaluated. The value of NCLOC peaked at 50 with the histogram before the peak resembles linear straight line and after the peak resembles inverse polynomial line. The value of approximation of both lines are shown in the Fig.1.

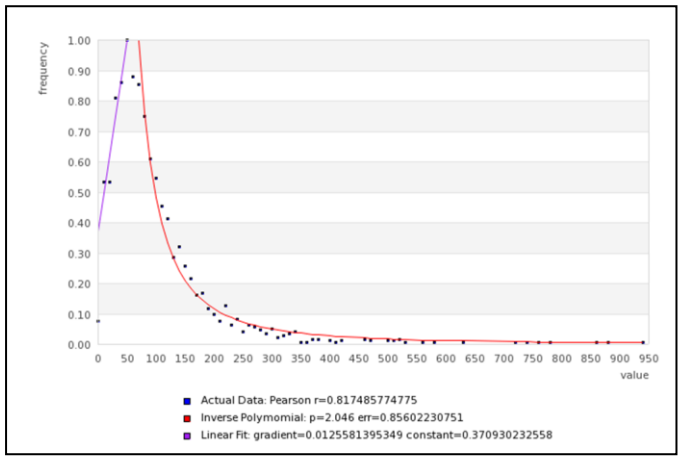


Figure 1. Histogram of Classes vs. NCLOC

If LOC_Q is defined as the normalized value of the quality of NCLOC, so the formula of LOC_Q are:

$$LOC_Q = 0.0125 \times NCLOC + 0.375 \text{ for } NCLOC \leq 50 \quad (1)$$

$$LOC_Q = (NCLOC - 50)^{-2.046} \text{ for } NCLOC > 50 \quad (2)$$

Where:

LOC_Q = NCLOC Quality Value

NCLOC = NCLOC Value

Note: the value of constant in formula (1) is adjusted from 0.371 into 0.375 to achieve the maximum value of 1 at NCLOC = 50.

2) Number of Functions: Figure 2 shows the histogram of classes vs. functions of all OSS Projects being evaluated. The peak value is 4.83 (rounded up into 5). Similar to class vs. NCLOC, the values before the peak resembles a straight line and after the peak resembles an inverse polynomial line with the approximation of both lines shown in the Fig.2.

F_Q is defined as the normalized value of function's quality, it can be formulated as follows:

$$F_Q = 0.172 \times F + 0.171 \text{ for } F \leq 5 \quad (3)$$

$$F_Q = (F - 4.83)^{-2.739} \text{ for } F > 5 \quad (4)$$

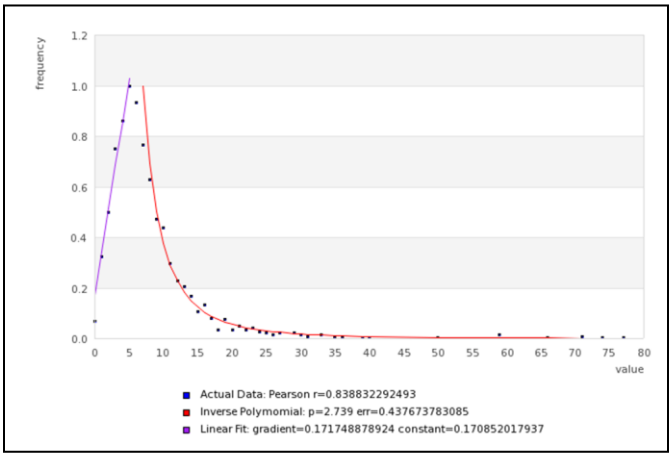


Figure 2. Histogram of Classes vs. Functions

Where:

F_Q = Function Quality Value

F = Number of Function

3) Cohesion: Cohesion is determined by the value of LCOM4. The ideal value is 1 which means that the class is highly cohesive. Higher value of LCOM4 indicates the degree of needed separation of classes into smaller classes.

$$LCOM4 \geq 1 \quad (5)$$

Where:

LCOM4 = Class Cohesion Value

4) Class Quality Formulation: Integrating all above measures into a single normalized value, the formulation of class quality or C_Q are:

$$C_Q = \frac{LOC_Q + F_Q}{2 \times LCOM4} \quad (6)$$

Where:

C_Q = Class Quality Value

LOC_Q = NCLOC Quality Value

F_Q = Function Quality Value

LCOM4 = Class Cohesion Value

B. Package Level Modularity

Package Quality or P_Q is the quality of individual package. Since in a single package there are many classes and there is no similarities found the the optimal number of classes in each package, so the Package Quality is determined by the average Class Quality or stated as:

$$P_Q = avg(C_Q) \quad (7)$$

Where:

P_Q = Package Quality Value

C_Q = Class Quality Value

C. System Level Modularity

S_A is a normalized value (with maximum value of 1) which determine the value of software architecture. The factors that influence this value are Package Cohesion (relationship among classes within package) and Package Coupling (relationship among classes from different packages). The principle used here is "*Maximize Cohesion and Minimize Coupling"* which becomes a widely known principle in building a good software system. The form of formulation is based on presentation titled "*Software Architecture Metrics*" by Ammar et. al [1], with the difference is that instead of using entropy approaches, this formulation is using the actual value of dependencies in determining the value of Package Cohesion and Package Coupling.

$$S_A = \sqrt{\frac{\sum_{i=1}^{d} C_{ii}^{\,2}}{\sum_{i=1}^{d}\sum_{j=1}^{d} C_{ij}^{\,2}}} \qquad (8)$$

Where:

Cii = Package Cohesion

Cij = Package Cohesion + Package Coupling

(if i=j is Package Cohesion,

if i ≠ j is Package Coupling)

d = number of package

D. Formulation of Modularity Index

Finally, the formulation of Modularity Index is the product of SA and the sum of all package quality in the software system as stated in the following formula:

$$M_I = S_A \times \sum_{i=1}^{j} P_{Qi} \qquad (9)$$

Where:

M_I = Modularity Index

S_A = Software Architecture Value

P_{Qi} = Package Quality of Package i

The proposed modularity index is a quality metrics will have the following properties:

- It has no upper bound: the value of modularity index increases as the number of module / package increases.
- The value of modularity index, especially the value of S_A depends on how the packages are coupled to each other. The limitation of connection of packages to only itself (package cohesion) or to only some dedicated packaged (e.g APIs, proxy, etc.) will improve the value of S_A.

V. Case Study: JFreeChart

JFreeChart is a free 100% Java chart library that makes it easy for developers to display professional quality charts in their applications (http://www.jfree.org/jfreechart) . This projects is one of the 59 OSS Projects used for modularity index formulation. For this case study, this project is chosen because:

- High S_A value (more than 0.7 since version 0.9.21)
- Relatively large number of packages (more than 30)

There are 52 versions available from the project's site, but only 33 are able to be analyzed using SONAR tool and being measured. The results are show in the following Fig.3.

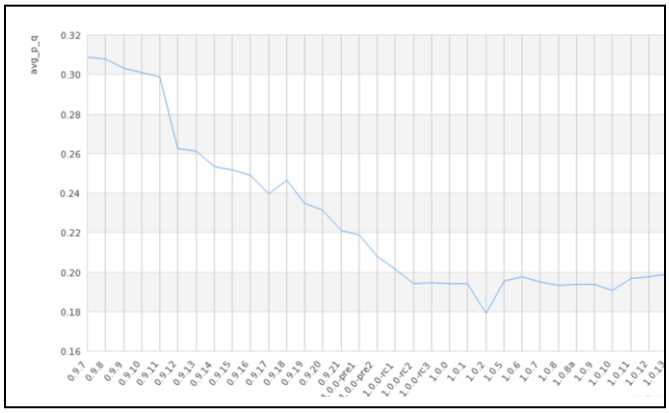


Figure 3. Average P_Q in 33 versions of JFreeChart

Fig. 3 above shows that the average package quality of the JFreeChart over 33 versions are decreasing consistently. This indicates the problem in the quality of each classes in each packages, such as:

- increasing size of NCLOC in each class.
- increasing number of functions in class.
- decreasing number of LCOM4 (Cohesion Metrics) in class.

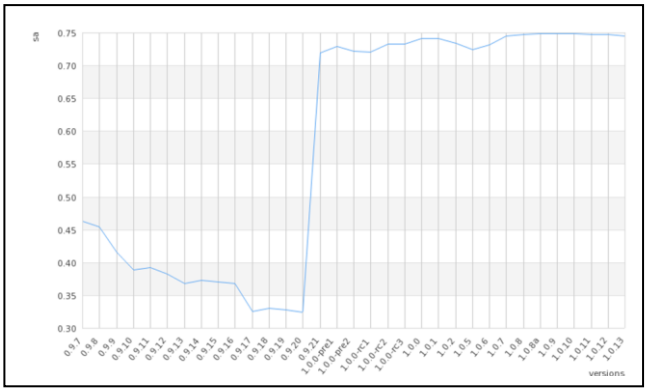


Figure 4. SA value in 33 versions of JFreeChart

Fig. 4 above shows that the structure of software architecture is improving. After consistent decrease in SA value in early versions of the system, there seems to be significant effort conducted before the release of version 1.0.0 started from version 0.9.21. The system from version 0.9.21 onward showing high number of SA.

The modularity index itself is shown in Fig. 5. The figure is showing improvement by the factor of two from early versions (until version 0.9.20) and late versions (version 1.0.5 onwards). There are significant jump in the value of modularity index from version 0.9.21 until version 1.0.2 indicating the period of major restructuring of the system before the release of milestone version 1.0.0.

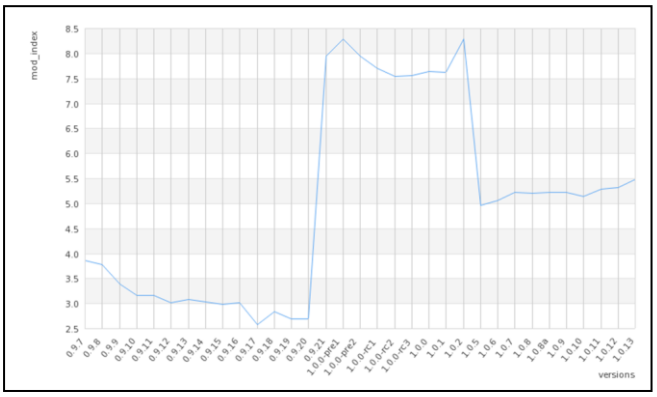


Figure 5. Modularity Index in 33 versions of JFreeChart

It can be seen from above case study that Modularity Index and its components (P_Q and S_A) are able to point the strength and potential problems in the development of JFreeChart OSS Projects. This information may give a valuable insight to the initiator and developers of the project in improving their project.

VI. Conclusion

Open Source Software (OSS) Projects are now gaining popularity and becoming one alternatives in developing software. Despite the the many success story of OSS Projects such as Apache, Mozilla, etc., the fact the many more of these projects that are failed needs are alarming. Some studies have identified that modularity is one of the key success factors of OSS Projects and authors believe that implementing modularity approach since early start of the project will increase the success of the project. This paper presents the first quantitative measure of modularity for Java-based OSS Projects called modularity index.

The formulation of modularity index are performed by analyzing the software metrics attributing to modularity of 59 Java-based OSS Projects from sourceforge.net which have been downloaded more than 100K times. By analyzing the similarity of these projects from class level, package level, and system level, the modularity index are formulated. As the validation of the software metrics, 33 out of 52 versions of JFreeChart OSS projects are analyzed using this metrics and the metrics are able to identify the strength and potential problems of the project.

Future study relating to this metrics involve further validation and integration into a framework called modularity framework in which the measurement of Modularity Frameworks will generate recommendations for improvement during OSS project's development. The integration of the software metrics into a web-based IDE will provide useful tool for project initiators and developers in improving their OSS Projects.

Acknowledgment

The authors would like to thank Maranatha Christian University (http://www.maranatha.edu) that provides funding for the research, and the Department of Computer Science and Electronics, Universitas Gadjah Mada (http://mkom.ugm.ac.id) that provides technical support for the research.

References

[1] Ammar H., Shereshevsky M., Mili A., Rabie W., Radetsky N. (2008), "Software architecture metrics", Seminar Presentation, Faculty of Information Science & Engineering, Management & Science University, Shah Alam, Malaysia, May 12, 2008. Available: http://www.docstoc.com/docs/6802629/Software-Architecture-Metrics

[2] Aruna M., M.P. Suguna Devi M.P, Deepa M. (2008), "Measuring the quality of software modularization using coupling-based structural metrics for an OSS system", Proceeding of the First International Conference on Emerging Trends in Engineering and Technology 2008

[3] Asundi J. (2007), "The need for effort estimation model for open source software projects", Proceeding of the Open Source Application Workspace: Fifth Workshop on Open Source Sofware Engineering 2007

[4] Bouktif S., Antoniol G., Merlo E. (2006), "A feedback based quality assessment to support open source software evolution: the GRASS case study", 22nd IEEE International Conference on Software Maintenance 2006, pp 155 - 165

[5] Cai Y., Huynh S. (2007), "An evolution model for software modularity assessment", Proceeding of the Fifth International Workshop on Software Qualty 2007 (WoSQ'07).

[6] Capiluppi A., Ramil J.F. (2004), "Studying the evolution of open source systems at different levels of granularity: two case studies", Proceeding on the 7th IEEE International Workshop of Principles of Software Evolution, 2004, pp 113 - 118.

[7] Capra E., Francalanci C., Merlo F. (2008), "An empirical study on the relationship among software design quality, development effort, and governance in open source projects", IEEE Transactions on Software Engineering Vol. 34, No. 6, Nov/Dec 2008, pp 765 – 782.

[8] Chidamber S.R., Kemerer C.F. (1994), "Metrics suite for object oriented design", IEEE Transaction on Software Engineering, Vol. 20 No. 6 June 1994, pp 476 – 493.

[9] Christley S., Madey G. (2007), "Analysis of activity in the open source software development community", Proceeding of the 40th IEEE Annual Hawaii International Conference on System Sciences, 2007, pp 166b.

[10] Crowston K., Wei K., Li Q., Howison J. (2006), "Core and periphery in free / libre and open source software team communications", Proceeding of the 39th IEEE Hawaii International Conference on System Sciences 2006

[11] DeKoenigsberg G. (2008), "How successful open source projects work, and how and why to introduce students to the open source world", 21st IEEE Conference on Software Engineering Education and Training, 2008, pp 274 – 276.

[12] Dinh-Trong T., Bieman J.M. (2004), "Open source software development: a case study of FreeBSD", Proceedings of the 10th IEEE International Symposium on Software Metrics, 2004, pp 96 - 105.

[13] Ellis H.J.C., Morelli R.A. , Lanerolle T.R., Damon J., Raye J. (2007), "Can humanitarian open-source software development draw new sudents to CS?", Proceeding of the 38th SIGCSE Technical Symposium on Computer Science Education 2007, pp 551 – 555.

[14] Emanuel A.W.R, Wardoyo R., Istiyanto J.E., Mustofa K. (2011), "Statistical analysis on software metrics affecting modularity in open source software", International Journal of Computer Science and Information Technology (IJCSIT), Vol. 3, No. 3, June 2011, pp 105 - 118

[15] Emanuel A.W.R, Wardoyo R., Istiyanto J.E., Mustofa K. (2010), "Success rules of OSS projects using datamining 3-itemset association rule", International Journal of Computer Science Issue (IJCSI), Vol. 7 Issue 6 Nov. 2010, pp 71 – 80.

[16] Emanuel A.W.R, Wardoyo R., Istiyanto J.E., Mustofa K. (2010), "Success factors of OSS projects from sourceforge using datamining association rule", Proceeding of 2010 International Conference on Distributed Frameworks for Multimedia Applications (DFmA) 2010, pp 141 - 148

[17] Gurbani V.K., Garvert A., Herbsleb J.D. (2005), "A case study of open source tools and practices in commercial setting", Proceeding of the fifth Workshop on Open Source Software Engineering 2005, pp 1 - 6.

[18] Hahsler M. (2005), "A quantitative study of the adoption of design patterns by open source software developers", Chapter V of Free / Open Source Software Development by Stefan Koch, Idea Group Publishing, ISBN 1-59140-371-5, 2005, pp 103 – 123.

[19] Hitz M., Montazeri B. (1995), "Measuring coupling and cohesion in object-oriented systems", Proceeding International Symposium on Applied Corporate Computing, Oct. 25-27 1995, Monterrey, Mexico, 75-76, 197, 78-84

[20] Lawrie T., Gacek C. (2002), "Issues of dependability in open source software development", Software Engineering Notes vol 27 no 3 of ACM Sigsoft. May 2002. Pp 34 -37

[21] Lee Y., Yang Y., Chang K.H. (2007), "Metrics & evolution in open source software", Proceeding on Seventh International Conference on Quality Software – 2007

[22] Li P.L., Herbsleb J., Shaw M. (2005), " Finding predictors of field defects for open source software systems in commonly available data sources: a case study of OpenBSD", Proceeding of 11th IEEE International Software Metrics Symposium, 2005, 32.

[23] McCabe T. (1976), "A complexity measure", IEEE Transactions On Software Engineering, Vol. Se-2, No. 4, December 1976, pp. 308-320.

[24] Melton H., Tempero E. (2007), "Toward assessing modularity", Proceeding of the First International Workshop on Assessment of Contemporary Modularization Techniques 2007 (ACoM'07)

[25] Meyer B., Oriol M., & Schoeller B. (2009), "Software engineering: lecture 17-18: estimation techniques and software metrics", Chair of Software Engineering Website, available: http://se.inf.ethz.ch/teaching/2008-S/se-0204/slides/15-Estimation-and-metrics-1-6x.pdf , accessed: 18 January 2009

[26] Mockus A., Fielding R.T., Herbsleb J.(2002), "Two case studies of open source software development: apache and mozilla", ACM Transaction on Software Engineering and Methodology Vol. II No. 3, Juli 2002, 309 – 346

[27] Mockus A, Fielding R.T, Herbsleb J.(2000), "A case study of open source software development: the apache server", ACM ICSE, 2000, 263 – 272

[28] Munelly J., Fritsch S., Clarke S. (2007). "An aspect-oriented approach to the modularisation of context". Proceedings of the Fifth Annual IEEE International Conference on Pervasive Computing and Communication (PerCom'07)

[29] Nakagawa E.Y, de Sousa E.P.M., de Britto Murata K. (2008), "Software architecture relevance in open source software evolution: a case study", Annual IEEE International Computer Software and Application Conference, 2008, pp 1234 – 1239.

[30] Paech B, Reuschenbach B (2006), "Open source requirements engineering", Proceeding of 14th IEEE International Requirement Engineering Conference: 257 - 262

[31] Raymond E.S. (2000), "The cathedral and the bazaar", version 3, Thyrsus Enterprises (http://www.tuxedo.org/~esr/), 2000.

[32] Spaeth S., Stuermer M. (2007), "Sampling in open source development: the case for using the debian GNU/linux dstribution", Proceedings of the 40th IEEE Hawaii International Conference on System Sciences, 2007, pp 166a.

[33] Stallman R. (1992), "Why software should be free", GNU Websites, 24 April 1992, Available: http://www.gnu.org/philosophy/shouldbefree.html

[34] Stewart K. J., Darcy D.P., Daniel S.L. (2005), "Observations on patterns of development in open source software projects", Proceeding on the fifth Workshop on Open Source Software Engineering 2005, pp 1 – 5.

[35] von Krogh G., Spaeth S., Haefliger S. (2005), "Knowledge reuse in open source software: an exploratory study of 15 open source projects", Proceeding of 38th Hawaii International Conference on System Sciences, 2005, pp. 198b

[36] Wang Y., Shao J. (2003), "Measurement of the cognitive functional complexity of software", Proceedings of the Second IEEE International Conference on Cognitive Informatics 2003 (ICCI'03).

[37] Zhou F., Davis J. (2008), "A model of bug dynamics for open source software", The Second IEEE International Conference on Secure System Integration and Reliability Improvement 2008, pp 185 - 186.

AUTHORS PROFILE

Andi Wahju Rahardjo Emanuel is a Full Time Lecturer at the Bachelor Informatics Program, Faculty of Information Technology, Maranatha Christian University in Bandung, Indonesia. He is graduated as BSEE in Purdue University, Indiana, USA in 1996 and MSSE in The University of Melbourne in 2001. He is currently taking his Doctoral Program at the Department of Computer Science and Electronics, Gadjah Mada University in Yogyakarta, Indonesia

Retantyo Wardoyo is an Associate Professor at the Department of Computer Science and Electronics, Universitas Gadjah Mada in Yogyakarta, Indonesia. He is graduated as Bachelor of Mathematics in Gadjah Mada University, Indonesia . He received his M.Sc in Computer Science in University of Manchester, UK and received his PhD in Computation in University of Manchester Institute of Science and Technology, UK.

Jazi Eko Istiyanto is a Professor and Head of the Department of Computer Science and Electronics, Universitas Gadjah Mada in Yogyakarta, Indonesia. He is graduated as Bachelor of Physics in Gadjah Mada University, Indonesia. He gets his Postgraduate Diploma (Computer Programming and Microprocessor), M.Sc (Computer Science) and PhD (Electronic System Engineering) at University of Essex, UK.

Khabib Mustofa is an Assistant Professor at the Department of Computer Science and Electronics, Universitas Gadjah Mada in Yogyakarta, Indonesia. He is graduated as Bachelor of Computer and Master of Computer at Gadjah Mada University, Indonesia. He receives his Dr. Tech in Computer Science at The Vienna University of Technology, Austria.

Survey of Nearest Neighbor Condensing Techniques

MILOUD-AOUIDATE Amal (corres.Author)
Computer Sciences Department
University of Sciences and Technology Houari Boumediene, USTHB,
Algiers, ALGERIA

BABA-ALI Ahmed Riadh
Electronic Department
University of Sciences and Technology Houari Boumediene, USTHB,
Algiers, ALGERIA

***Abstract*—The nearest neighbor rule identifies the category of an unknown element according to its known nearest neighbors' categories. This technique is efficient in many fields as event recognition, text categorization and object recognition. Its prime advantage is its simplicity, but its main inconvenience is its computing complexity for large training sets.**
This drawback was dealt by the researchers' community as the problem of prototype selection.
Trying to solve this problem several techniques presented as condensing techniques were proposed. Condensing algorithms try to determine a significantly reduced set of prototypes keeping the performance of the 1-NN rule on this set close to the one reached on the complete training set. In this paper we present a survey of some condensing KNN techniques which are CNN, RNN, FCNN, Drop1-5, DEL, IKNN, TRKNN and CBP.
All these techniques can improve the efficiency in computation time. But these algorithms fail to prove the minimality of their resulting set. For this, one possibility is to hybridize them with other algorithms, called modern heuristics or metaheuristics, which, themselves, can improve the solution. The metaheuristics that have proven results in the selection of attributes are principally genetic algorithms and tabu search. We will also shed light in this paper on some recent techniques focusing on this template.

Keywords- Nearest neighbor (NN); kNN; Prototype selection; Condensed NN; Reduced NN; Condensing; Genetic algorithms; Tabu search.

I. INTRODUCTION

The K-nearest neighbor classification rule (KNN) proposed by T. M. Cover and P. E. Hart [4], is a powerful classification method that allows an almost infallible classification of an unknown prototype through a set of training prototypes. It is widely used in pattern recognition [20] [18], text categorization [10] [6], object recognition [8] and event recognition [23] applications.

An inevitable consequence of large sets of prototypes is the computational time implied by this research problem. The databases, used in some areas such as intrusion detection, are constantly and dynamically updated. This constitutes one of the main inconveniences of the KNN rule. Another important inconvenience comes from the fact that the training prototypes can contain noisy or mislabeled prototypes that may affect the results and distort them. The scientific community has tackled these problems and proposed a selection of prototypes which could modify an initial set of prototypes by reducing its size in order to improve the classification performance.

II. PROTOTYPE SELECTION

Prototype selection is the process of finding representative patterns from the data, which can help in reducing these data. This problem is classified as an NP-hard problem by many researchers [1] [25], because there is no polynomial algorithm allowing for the solution. The existing algorithms can just give acceptable solutions.

Like many other combinatorial problems, the prototype selection (PS) would require an exhaustive search to obtain optimal solutions in the general case. This has led some researchers to consider the problem of PS as a combinatorial optimization problem and use general techniques which are known for their good results in similar situations.

Heuristics, especially the nearest neighbor algorithm, and metaheuristics, especially genetic algorithms (GA) and tabu search (TS) have been proposed to solve this problem.

III. IMPROVING PROTOTYPE SELECTION

Condensing algorithms try to find a significant reduction of all prototypes so that the 1-NN classification gives results as close as possible to those obtained using all the original prototypes **[7]**. The problem that arises when this approach is used is that it cannot provide a proof about the resulting sets minimality. To correct this slight defect, the modern heuristics or metaheuristics came to complete them. The metaheuristics are strategies that guide the search towards an optimal solution. These techniques are designed to explore the search space efficiently in order to determine solutions (almost) optimal. They may contain mechanisms to avoid blocking in areas of space research.

Two types of metaheuristics have been successful in their hybridization with the traditional KNN: genetic algorithms and tabu search.

Genetic algorithms (GA) are an optimization technique guided by the principles of natural evolution and genetics, with a high presence of implicit parallelism. These algorithms perform a search in complex, large, and multimodal landscapes, and provide solutions for the quasi-optimal objective function.

The tabu search (TS) is a method of dynamic neighborhood, which selects, at each iteration, the best solution of the first local optimum by finding the best neighbor.

IV. HEURISTIC METHODS BASED ON THE CONDENSING K-NN RULE

In 1968 Hart [17] was the first to propose a method reducing the size of stored data for the nearest neighbor decision. This method is called "*The Condensed Nearest Neighbor Rule*" (CNN). The new idea about this rule compared to the traditional KNN is the process of selecting a subset TCNN from the initial training set TNN. TCNN must be as effective as TNN in the classification of unknown patterns. Actually the CNN rule minimizes the number of models stored, keeping only a subset of training data for classification, and employing a technique of low absorption. The basic idea is to look for very similar training models, and those that do not add additional information and eliminate them.

This rule presents a case where TCNN consistency is not achieved, when TNN is already the minimum set. In this case TCNN will be equal to TNN, if this happens, the algorithm will end if there are two equal models of different classes, but TCNN must classify patterns correctly.

The technique that corrects this case of inconsistency is the "*Reduced Nearest Neighbor rule"* (RNN) introduced by Gates [11]. This rule is an extension of the CNN rule, and like CNN, RNN reduces TNN.

The RNN algorithm starts at TRNN=TCNN and removes every instance from TRNN if this deletion does not cause a misclassification of another instance in TNN by the remaining instances in TRNN. From the perspective of computing, it is more expensive than the rule proposed by Hart, but it will always produce a subset of CNN, and thus will be less expensive in terms of computing and storage at the classification stage.

Angiulli introduced the "*Fast Condensed Nearest Neighbor rule"* (FCNN) [2] a scalable algorithm on large multidimensional data sets used to create subsets serving as consistent training sets based on the nearest neighbor decision rule. This algorithm allows selecting points very close to the border decision. It is independent of the order, and has low quadratic complexity.

The FCNN rule initializes the consistent subset S with a starting element from each class label of the training set. The starting elements that the rule uses are particularly the class's centroids in the training set. The algorithm is incremental. At each iteration the resulting set is increased till the stop condition is reached.

Wilson and Martinez suggested [21] a series of six algorithms for sets reduction based on the kNN algorithm where each algorithm improves the previous one. The first reduction technique presented was the DROP1 which constitutes the basic framework on which were built the five other techniques**.** The DROP1 represents an improvement of the RNN rule, which verifies the accuracy of the resulting set instead of the initial set T. This algorithm is based on the following rule: an instance P is removed only if at least some of its associates (neighbors from same class) in S can be classified correctly without P.

The first proposal causes a problem when the noisy instances, which are typically associated with a different class, cover only a small portion of the input space. DROP2 tries to solve this problem by considering the effect of removing an instance on all instances of the initial training T rather than S. For this purpose the rule of DROP1 has been improved to eliminate P if at least an acceptable number of its associates in T can be classified correctly without P.

DROP2 sorts S in an attempt to remove the central points before the border points (points which are near from the decision boundaries), however noisy instances can also be on borders, which can cause a change in the removal order. And even if a noisy instance is central attempting to remove it could eliminate border points that should be maintained. Hence DROP3 uses a noise filtering before sorting the instances of S. This is done using a rule that eliminates any instance misclassified by its k nearest neighbors. Except that DROP3 can sometimes remove an overly large number of instances.

DROP4 improves DROP3 rule and provides that an instance is removed only if:

1. It is misclassified by its k nearest neighbors, and
2. Its removal does not affect the classification of other instances

DROP5 upgrades DROP2 by proposing that the instances are considered beginning from the ones closest to the nearest enemy (an enemy is the nearest neighbor of an instance with a different class) and proceeding to outside.

The latest algorithm proposed by Wilson and Martinez [22] was the DEL which is similar to DROP3, except that it uses the length coding heuristic for deciding whether an instance can be removed or not. In DEL an instance is removed only if:

1. It is misclassified by its k nearest neighbors, and
2. The removal of the instance does not increase the cost of length encoding

Wu, Ianakiev and Govindraju proposed an "*Improved K*-Nearest Neighbor Classification" [24]. A solution to increase the speed of traditional kNN classification while maintaining its level of accuracy by suggesting two building techniques. The suggested IKNN algorithm is based on iterative elimination of models with high attraction capacity.

In the first technique called the model condensing, the authors suggested that all classes have the same probability, and that the training set Ω is initially created by the extraction of vector elements from a large set of images of interest, where each class has an equal representation. This set Ω is refined iteratively.

In the second technique, which is the *pre-processing*, an unknown pattern is compared to a prototype in two stages. In the first stage a rapid evaluation of the potential match is made. The prototypes that fail in the first match are not included in the second. Then, for a complete match, in the

second stage the norms difference of the prototype and the test pattern must be less than a predetermined threshold.

The observation that large sets of data had computational requirements which could be prohibitive to classify models using the kNN, has led Fayed and Atia [9] to propose the TRKNN, a way to lessen this problem through a condensation approach. The aim of their approach is to eliminate the reasons that makes load the calculation and does not contribute to improve the classification.

This approach consists in rejecting the prototypes that are far from the limits and have just a little influence on the KNN classification. To achieve this, the authors first introduced the concept of chain of nearest neighbors which is a sequence of the nearest neighbors from alternating classes.

This technique proposes that if the distances of a given model decrease in value, this model is considered to be probably an internal point and can be discarded, whereas if the distances do not decrease too much, then this point probably varies around the limit of classification and it is maintained.

In a more recent paper [14] the authors introduced a new approach *"The Class Boundary Preserving Algorithm"* (CBP), a multi-step method for pruning the training set. The proposed method aims at preserving instances that are close to the borders of classes. Because, according to the authors, these instances can provide most of the necessary information to properly describe the underlying distribution. On the other hand instances distant from limits are considered redundant by the authors.

The innovation in this approach is the procedure used to divide the training set into two subsets, X_B containing the instances near the surface of decision, and X_{NB} containing the internal samples. And because there is a noticeable difference in the importance of information held by these two sets, two different reduction processes were applied on both.

The algorithm considers an initial set of n instances related to a set of labels, and involves four main steps.

The first step deals with the smoothing of class boundaries, the second allows to distinguish between border and non-border instances. The third step has to do with the pruning of border instances and the last step is the clustering of non-border instances.

To implement this idea, the authors have used the *Mean Shift Clustering algorithm* (MSC) which converges to the points of maximum density to determine the cluster centers of the distribution. Then to make the maximum number of neighbors in obtained clusters, the authors applied a merger process which verifies the calculated clusters centers labels, and if the clusters of nearest neighbors share the same label, then the centers are merged.

A. Comparison

After examining some articles that have tackled the reduction of sets based on the condensing KNN, we consider useful to raise a comparative board grouping the idea, the advantages, and disadvantages of each algorithm in the cited articles.

TABLE I. COMPARISON OF NEAREST NEIGHBOR CONDENSING TECHNIQUES

CNN	**Idea**	Remove Training models that are very similar and those that do not add additional information to the classification
	Advantages	- Improves search time and memory requirements - Reduces the size of training data.
	disadvantages	- CNN is order dependent, then it is unlikely that it removes border points - If the initial set is minimal, that causes an inconsistency in the resulting set when the program is stopped. - There is no guarantee that the resulting set is minimal.
RNN	**Idea**	Initially the resulting set is equal to the initial set, and then each instance that does not cause a wrong classification of another instance in the initial set is removed from the resulting set.
	Advantages	- Reduces the size of training data and eliminates models - Improves the search time and memory requirements.
	Disadvantages	- High computational cost - Time consuming - Its consistency depends on the consistency of the resulting set of CNN
FCNN	**Idea**	Select points very close to the decision boundary
	Advantages	- Average efficiency of 96.01% for an average number of iterations of about 69 iterations - Has a smaller complexity than CNN - Good rate of condensation - Independent of the order
	Disadvantages	- Requires a large number of iterations
DROP1	**Idea**	Initially the resulting set is equal to the initial set, and then an instance is removed only if at least some of his associates in the resulting set can be ordered without it.
	Advantages	- Reduces the size of the training data and eliminates instances - Constitutes a basis on which are built the rest of the DROP algorithms and DEL - Noise do not degrade accuracy
	Disadvantages	- Checks the consistency of the resulting set instead of the initial set - Low accuracy - Cannot use information from previously eliminated instances
DROP2	**Idea**	Improves DROP1 and eliminates an instance when at least a good number of its associates in the initial set can be classified without it.
	Advantages	- Checks the consistency of the initial set rather than the final set

		- Reduces the size of training data and eliminates instances - Reach a higher accuracy than KNN for noisy instances - Good storage reduction happens to 1 / 6 of the original set - Storage requirements lower than DROP 1, 4, 5
	Disadvantages	- Attempts to remove the central points before the borderer, which overlooks the border noisy instances - The attempt to eliminate a noisy central instance can remove border points that should be maintained. - For very large data sets DROP2 does not eliminate enough points
DROP3	**Idea**	Improves DROP2 by eliminating instances misclassified by their neighbors
	Advantages	- Is based on the classification of the instance itself to remove it - Reduces the size of training data and eliminates instances - Reach a higher accuracy than the traditional KNN in case of noisy instances - Very good reduction of storage that comes to 12% of the original set - Storage requirements lower than DROP1, 4, 5
	Disadvantages	- Can sometimes remove a too large number of instances
DEL	**Idea**	Improves DROP3 and eliminates an instance if: 1. It is misclassified by its k nearest neighbors, and 2. The removal of the instance does not increase the cost of length encoding
	Advantages	- Achieved higher accuracy than the traditional KNN in case of noisy instances - Reduces the size of training data and eliminates instances
	Disadvantages	- Accuracy less than that of DROP2-5 - Storage requirements higher than those of DROP2-5
IKNN	**Idea**	Remove iteratively models exhibitors of high capacity of attraction
	Advantages	- Reduces the data set by keeping the prototypes that are useful - The preprocessing allows a significant saving in computation time - The classifier shows a slight improvement in accuracy compared to traditional kNN - Reduces the size of the models maintaining the same level of accuracy
	Disadvantages	- Classes must have same probability in the training set - The work was built on intuition and not on a mathematical framework citing that the norm is an intrinsic characteristic - No theoretical proof that preprocessing guaranteed to filter the relevant prototypes or to maintain the accuracy - A test pattern is a distorted version of a prototype when the difference in standard is below a threshold associates to this prototype
TRKNN	**Idea**	Eliminate the patterns that are a burden on computing and does not contribute to improve the classification
	Advantages	- Reduces the size of the models without sacrificing accuracy - TRKNN is up to 3 times faster than IKNN and up to 4 times than DROP2
	Disadvantages	- Average condensation rate (35%) - Average level of accuracy
CBP	**Idea**	Preserve instances that are close to class boundaries
	Advantages	- Combine the selection and the abstraction - Applies to each case an appropriate procedure of condensation - The filtering Phase is based on the classification of the instance itself - Use the geometric characteristics of the distribution - Provides an overview of the distribution of samples - Reduces the size of the training set and the number of representatives - Good rate of condensation
	Disadvantages	- The filtering algorithm can sometimes remove a very large number of instances - Relatively high computation time

These approaches which are based on traditional KNN, select all the prototypes from training samples of initial set, by adding or cutting, with aiming at preserving the performance of classification with the use of heuristic methods. However, none of these methods can prove the minimality of the result set.

V. METAHEURISTICS: IMPROVING KNN

The basic concepts of metaheuristics can be described in the abstract, without requiring a specific problem [15]. These algorithms can therefore use heuristics, which in turn reflect the specificity of the problem treated, except that these heuristics are controlled by a higher level strategy.

In this article we present some improvements of classical kNN algorithms through the use of some metaheuristics such as genetic algorithms and tabu search.

R. Gil-Pita and X. Yao proposed [12] three improvements of the k-nearest neighbor using genetic algorithms: the use of an objective function based on mean square error, the implementation of a clustered crossover, and a fast smart mutation scheme.

In the first proposal they make use of a novel objective function based on the MSE function. They consider the kNN as a system with C outputs, so that each output is calculated using a defined equation. Therefore, the C outputs of the kNN

system are approximations of the posterior probabilities of the data.

In the second proposal the authors suggest a new crossover scheme for the GA, denominated *clustered crossover* (CC), in order to improve the determination of the best subset of the training set that minimizes the selected objective function by considering possible to determine the relationship between the different bits of the bit stream.

In the third proposal R. Gil-Pita and X. Yao describe the application of a mutation scheme that allows selecting the best gene or group of genes to be changed, taking into account the variations of the objective function with respect to each gene for a given set. They design a fast method for evaluating the error variation when each gene is changed, and they propose a mutation strategy based on these variations of the objective function. The authors denominate this mutation scheme as *fast smart mutation* (FSM) [13], as it allows increasing the effectiveness of the mutation stage in the genetic algorithm.

Compared to the use of a classic kNN classifier, obtained results demonstrate the good accuracy of the proposed GA-based technique.

The authors of "*A hybrid classification method of k-nearest neighbor, Bayesian methods and genetic algorithm*" [16] introduced a hybrid technique including KNN, genetic algorithms and Bayesian method. This approach consists of eight steps that begin with the application of the Bayesian algorithm and then the generation of new data using the genetic algorithm and the application of the K-nearest neighbor's method. And finally, several iterations of these algorithms occur orchestrated by the genetic mechanism. According to the authors, the method is useful on data sets that have a small amount of data. It generates an unlimited number of data that have similar characteristics to the original data, and improves these data according to the proposed algorithm.

This method shows better classification performance with respect to classic methods such as expectation maximization algorithm [5] used to develop it. It is intended, according to its authors, to hardware solutions with low cost based on clustering, to noisy data classification, and to classification in the data sets with little data.

To overcome the limitations of kNN, an improved version of KNN, "*Genetic KNN"* (GKNN), was proposed by Suguna and Thanushkodi [19]. A genetic algorithm is combined with the k-nearest neighbor algorithm (KNN). In the proposed method, using the genetic algorithm, k-number of samples are selected for each iteration and the accuracy of the classification is calculated as fitness. The greater accuracy is recorded each time. Thus, it is not necessary to calculate the similarities between all samples, or to consider the weight of the category.

The performance of GKNN classifier was compared with traditional KNN. The experiments and results show that the GKNN not only reduces the complexity of the KNN, but it improves the classification accuracy.

Ceveron and Ferri [3] proposed a method for the prototypes selection for the nearest neighbor rule which aims to obtain an optimal or close to optimal solution by presenting a new approach based on tabu search. The particularity of the proposed tabu search is that it uses the equations from objective genetic algorithms. In this approach all possible subsets of prototypes constitute the space of solutions. Possible moves from a particular subset consist of adding or removing each of the n initial prototypes. The attribute used to declare the movement taboo is the prototype that is added or deleted.

It is worth mentioning that the results obtained with TS consistently improved the classical condensing techniques previously published in the literature using the Iris database.

Wu's article [22] provides a method of selecting items based on the tabu search algorithm and KNN. First, the KNN algorithm is used to generate the initial solution needed for the tabu search. Then, the tabu algorithm is applied to obtain an optimal subset of items. The KNN algorithm uses the relevance of the elements to eliminate those redundant in large networks data, and the subset obtained is the initial solution of the tabu search algorithm.

This algorithm was tested via creating intrusion detection model. The tests show that by using the feature selection method proposed in this paper, the detection performance of the intrusion detection system is effectively improved without compromising detection accuracy, and the detection time and accuracy performance of the system is much better than that of the current feature selection methods. As a result the method proposed in this paper has proved to be effective and feasible.

VI. CONCLUSION

In this paper we tried to present and compare sets reduction techniques based on the principle of nearest neighbor. These techniques are of type "condensing". Both are improvements compared to basic KNN. These improvements have been proposed by the authors to reduce the training set to gain on speed and space efficiencies. To ensure the minimality of this training set we presented some recent proposals using metaheuristics to check the optimality of the resulting set of some KNN reduction techniques.

Note that each technique is very effective in a specific area and in special circumstances.

REFERENCES

[1] Chang Bao Rong, Naghibzadeh Mahmoud, Czarnowski Ireneusz, Danesh Malihe and Danesh Mohaddesh, "Data Clustering Based on an Efficient Hybrid of K-Harmonic Means, PSO and GA", Transactions of Computational Collective Intelligence, vol. IV, pp. 125-140, Springer, 2011.

[2] Fabrizio Angiulli, "Fast condensed nearest neighbor rule", Technical report, Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005.

[3] Vicente Cerveron and Francesc J. Ferri, "Another move toward the minimum consistent subset: A tabu search approach to the condensed nearest neighbor rule", IEEE Transactions on Systems, Man, and Cybernetics, vol. 31, pp. 408–413, 2010.

[4] T. M. Cover and P. E. Hart, "nearest neighbor pattern classification", IEEE Transaction on Information Theory, vol. 13, pp. 21–27, Jan 1967.

[5] F. Dallaert, "The expectation maximization algorithm", Technical report, College of Computing, Georgia Institute of Technology, February 2002.

[6] E. M. Elnahrawy, "Log based chat room monitoring using text categorization: A comparative study", University of Maryland.

[7] Pavel Paclik, Elzbieta Pekalska and Robert P.W. Duin, "Prototype selection for

dissimilarity-based classifiers", Pattern Recognition, vol. 39, pp. 189–208, 2006.

[8] F. Bajramovic, Frank Mattern, Nicholas Butko, and Joachim Denzler "A comparison of nearest neighbor search algorithms for generic object recognition", LNCS 4179, ACIVS 2006, pp. 1186–1197.

[9] Hatem A. Fayed and Amir F. Atiya, "A novel template reduction approach for the K-nearest neighbor method", IEEE Transactions on Neural Networks, vol. 20 (5), pp. 890–896, May 2009.

[10] O. Kirmemis and G. Toker, "Text categorization using k nearest neighbor classification", Survey paper, Middle East Technical University.

[11] G.Gates, "The reduced nearest neighbor rule". IEEE Transactions on Information Theory, vol. 18, pp. 431–433, 1972.

[12] R. Gil-Pita and X. Yao, "Using a genetic algorithm for editing k-nearest neighbor classifiers", IDEAL 2007, LNCS, vol. 4881, pp. 1141-1150, 2007.

[13] X Gil-Pita, R. Yao, "Evolving edited k-nearest neighbor classifiers", International Journal of Neural Systems, vol. 18 (6), pp. 459–467, Dec 2008.

[14] Q.H. Wu, K. Nikolaidis and J.Y. Goulermas, "A class boundary preserving algorithm for data condensation", Pattern Recognition, vol. 44, pp. 704-715, 2011.

[15] Sean Luke, "Essentials of Metaheuristics". Lulu, 2009.

[16] Mutlu Avci Mehmet Aci and Cigdem Inan, "A hybrid classification method of k nearest neighbor, bayesian methods and genetic algorithm", Expert Systems with Applications, vol. 37, pp. 5061-5067, 2010.

[17] P. Hart, "The condensed nearest neighbor rule", IEEE Transactions on Information Theory, vol. 14, pp. 515–516, 1968.

[18] Y. Wu Shizen, "An algorithm for remote sensing image classification based on artificial immune b-cell network", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXVII, Part B6b, pp. 107-111, Beijing 2008.

[19] Suguna and Dr. K. Thanushkodi, "An improved k-nearest neighbor classification using genetic algorithm", International Journal of Computer Science Issues, vol. 7, pp. 18–21, 2010.

[20] V.Vaidehi, S. Vasuhi, R.Kayalvizhi, K.Mariammal, Raghuraman.M.B, "Person authentication using face recognition", Proceedings of the World Congress on Engineering and Computer Science, 2008.

[21] Randall Wilson and Tony R. Martinez, "Reduction techniques for instance-based learning algorithms", Machine Learning, vol. 38 (3), pp. 257–286, 2000.

[22] Tao Ran, Wu Jian-guang and Li Zhi-Yong, "An improving tabu search algorithm for intrusion detection", IEEE Computer Society, pp. 435–439, 2011.

[23] Y. Yang and T. Ault, "Improving text categorization methods for event tracking", Carnegie Mellon University.

[24] Krasimir G. Iankiev Yingquan Wu and Venu Govindaraju, "Improved k-nearest neighbor classification", Pattern recognition, vol. 35, pp. 2311–2318, 2002.

[25] A. V. Zukhba, "Np-completeness of the problem of prototype selection in the nearest neighbor method", Pattern Recognition and Image Analysis, vol. 20, pp. 484-494, 2010.

Concurrent Edge Prevision and Rear Edge Pruning Approach for Frequent Closed Itemset Mining

Anurag Choubey[1]
Dean Academic, Technocrats Institute of Technology, RGPV, Bhopal (M.P.), India

Dr. Ravindra Patel[2]
Reader & Head, Dept. of Computer Application, UIT-RGPV, Bhopal (M.P.), India

Dr. J.L. Rana[3]
Former Professor & Head, Dept. of CSE & IT MANIT, Bhopal (M.P.), India

***Abstract*- Past observations have shown that a frequent item set mining algorithm are purported to mine the closed ones because the finish provides a compact and a whole progress set and higher potency. Anyhow, the newest closed item set mining algorithms works with candidate maintenance combined with check paradigm that is pricey in runtime yet as space usage when support threshold is a smaller amount or the item sets gets long. Here, we show, CEG&REP that could be a capable algorithm used for mining closed sequences while not candidate. It implements a completely unique sequence finality verification model by constructing a Graph structure that build by an approach labeled "Concurrent Edge Prevision and Rear Edge Pruning" briefly will refer as CEG&REP. a whole observation having sparse and dense real-life knowledge sets proved that CEG&REP performs bigger compared to older algorithms because it takes low memory and is quicker than any algorithms those cited in literature frequently.**

Keywords- Data mining; Closed Itemsets; Pattern Mining; sequence length; graph structure.

I. INTRODUCTION

Sequential item set mining, is an important task, having many applications with market, customer and web log analysis, item set discovery in protein sequences. Capable mining techniques are being observed extensively, including the general sequential item set mining [1, 2, 3, 4, 5, 6], constraint-based sequential item set mining [7, 8, 9], frequent episode mining [10], cyclic association rule mining [11], temporal relation mining [12], partial periodic pattern mining [13], and long sequential item set mining [14]. Recently it's quite convincing that for mining frequent item sets, one should mine all the closed ones as the end leads to compact and complete result set having high efficiency [15, 16, 17, 18], unlike mining frequent item sets, there are less methods for mining closed sequential item sets. This is because of intensity of the problem and CloSpan is the only variety of algorithm [17], similar to the frequent closed item set mining algorithms, it follows a candidate maintenance-and-test paradigm, as it maintains a set of readily mined closed sequence candidates used to prune search space and verify whether a recently found frequent sequence is to be closed or not. Unluckily, a closed item set mining algorithm under this paradigm has bad scalability in the number of frequent closed item sets as many frequent closed item sets (or just candidates) consume memory and leading to high search space for the closure checking of recent item sets, which happens when the support threshold is less or the item sets gets long.

Finding a way to mine frequent closed sequences without the help of candidate maintenance seems to be difficult. Here, we show a solution leading to an algorithm, CEG&REP, which can mine efficiently all the sets of frequent closed sequences through a sequence graph protruding approach. In CEG&REP, we need not eye down on any historical frequent closed sequence for a new pattern's closure checking, leading to the proposal of Sequence graph edge pruning technique and other kinds of optimization techniques.

The observations display the performance of the CEG&REP to find closed frequent itemsets using Sequence Graph: The comparative study claims some interesting performance improvements over BIDE and other frequently cited algorithms.

In section II most frequently cited work and their limits explained. In section III the Dataset adoption and formulation explained. In section IV, introduction to CEG&REP and its utilization for Sequence Graph protruding explained. In section V, the algorithms used in CEG&REP described. In section V1, results gained from a comparative study briefed and fallowed by conclusion of the study.

II. RELATED WORK

The sequential item set mining problem was initiated by Agrawal and Srikant , and the same developed a filtered algorithm, GSP [2], basing on the Apriori property [19]. Since then, lots of sequential item set mining algorithms are being developed for efficiency. Some are, SPADE [4], PrefixSpan [5], and SPAM [6]. SPADE is on principle of vertical id-list format and it uses a lattice-theoretic method to decompose the search space into many tiny spaces, on the other hand PrefixSpan implements a horizontal format dataset representation and mines the sequential item sets with the pattern-growth paradigm: grow a prefix item set to attain longer sequential item sets on building and scanning its database. The SPADE and the PrefixSPan highly perform GSP. SPAM is a recent algorithm used for mining lengthy sequential item sets and implements a vertical bitmap representation. Its observations reveal, SPAM is better efficient in mining long

item sets compared to SPADE and PrefixSpan but, it still takes more space than SPADE and PrefixSpan. Since the frequent closed item set mining [15], many capable frequent closed item set mining algorithms are introduced, like A-Close [15], CLOSET [20], CHARM [16], and CLOSET+ [18]. Many such algorithms are to maintain the ready mined frequent closed item sets to attain item set closure checking. To decrease the memory usage and search space for item set closure checking, two algorithms, TFP [21] and CLOSET+2, implement a compact 2-level hash indexed result-tree structure to keep the readily mined frequent closed item set candidates. Some pruning methods and item set closure verifying methods, initiated the can be extended for optimizing the mining of closed sequential item sets also. CloSpan is a new algorithm used for mining frequent closed sequences [17]. It goes by the *candidate maintenance-and-test* method: initially create a set of closed sequence candidates stored in a hash indexed result-tree structure and do post-pruning on it. It requires some pruning techniques such as *Common Prefix* and *Backward Sub-Item set pruning* to prune the search space as CloSpan requires maintaining the set of closed sequence candidates, it consumes much memory leading to heavy search space for item set closure checking when there are more frequent closed sequences. Because of which, it does not scale well the number of frequent closed sequences. BIDE **[**26**]** is another closed pattern mining algorithm and ranked high in performance when compared to other algorithms discussed. Bide projects the sequences after projection it prunes the patterns that are subsets of current patterns if and only if subset and superset contains same support required. But this model is opting to projection and pruning in sequential manner. This sequential approach sometimes turns to expensive when sequence length is considerably high. In our earlier literature[27] we discussed some other interesting works published in recent literature.

Here, we bring Sequence Graph protruding that based on edge projection and pruning, an asymmetric parallel algorithm for finding the set of frequent closed sequences. The giving of this paper is:

(A) an improved sequence graph based idea is generated for mining closed sequences without candidate maintenance, termed as Concurrent Edge Prevision and Rear Edge Pruning (CEG&REP) based Sequence Graph Protruding for closed itemset mining. The Edge Projection is a forward approach grows till edge with required support is possible during that time the edges will be pruned. During this pruning process vertices of the edge that differs in support with next edge projected will be considered as closed itemset, also the sequence of vertices that connected by edges with similar support and no projection possible also be considered as closed itemset

(B) in the Edge Projection and pruning based Sequence Graph Protruding for closed itemset mining, we create a algorithms for Edge Prevision and Rear Edge Pruning

(C) The performance clearly signifies that proposed model has a very high capacity: it can be faster than an order of magnitude of CloSpan but uses order(s) of magnitude less memory in several cases. It has a good scalability to the database size. When compared to BIDE the model is proven as equivalent and efficient in an incremental way that proportional to increment in pattern length and data density.

III. DATASET ADOPTION AND FORMULATION

Item Sets I: A set of diverse elements by which the sequences generate.

$$I = \bigcup_{k=1}^{n} i_k$$

Note: 'I' is set of diverse elements

Sequence set 'S': A set of sequences, where each sequence contains elements each element 'e' belongs to 'I' and true for a function p(e). Sequence set can formulate as

$$s = \bigcup_{i=1}^{m} < e_i \mid (p(e_i), e_i \in I) >$$

Represents a sequence's' of items those belongs to set of distinct items 'I'.

'm': total ordered items.

$P(e_i)$: a transaction, where e_i usage is true for that transaction.

$$S = \bigcup_{j=1}^{t} s_j$$

S: represents set of sequences

't': represents total number of sequences and its value is volatile

s_j: is a sequence that belongs to S

Subsequence: a sequence s_p of sequence set 'S' is considered as subsequence of another sequence s_q of Sequence Set 'S' if all items in sequence S_p is belongs to s_q as an ordered list. This can be formulated as

$$(\bigcup_{i=1}^{n} s_{pi} \in s_q) \Rightarrow (s_p \subseteq s_q)$$

If

Then $\bigcup_{i=1}^{n} s_{pi} <: \bigcup_{j=1}^{m} s_{qj}$ where $s_p \in S \text{ and } s_q \in S$

Total Support 'ts' : occurrence count of a sequence as an ordered list in all sequences in sequence set 'S' can adopt as total support 'ts' of that sequence. Total support 'ts' of a sequence can determine by fallowing formulation.

$$f_{ts}(s_t) = \mid s_t <: s_p \ (\textit{for each } p = 1..\mid DB_S \mid) \mid$$

DB_S Is set of sequences

$f_{ts}(s_t)$: Represents the total support 'ts' of sequence s_t is the number of super sequences of s_t

Qualified support 'q_s': The resultant coefficient of total support divides by size of sequence database adopt as qualified support 'qs'. Qualified support can be found by using fallowing formulation.

$$f_{qs}(s_t) = \frac{f_{ts}(s_t)}{|DB_S|}$$

Sub-sequence and Super-sequence: A sequence is sub sequence for its next projected sequence if both sequences having same total support.

Super-sequence: A sequence is a super sequence for a sequence from which that projected, if both having same total support.

Sub-sequence and super-sequence can be formulated as

If $f_{ts}(s_t) \geq$ rs where 'rs' is required support threshold given by user

And $s_t \lessdot s_p$ *for any p value* where $f_{ts}(s_t) \cong f_{ts}(s_p)$

IV. Concurrent Edge Prevision and Rear Edge Pruning

A. Preprocess:

As a first stage of the proposal we perform dataset preprocessing and itemsets Database initialization. We find itemsets with single element, in parallel prunes itemsets with single element those contains total support less than required support.

B. Edge Prevision:

In this phase, we select all itemsets from given itemset database as input in parallel. Then we start projecting edges from each selected itemset to all possible elements. The first iteration includes the pruning process in parallel, from second iteration onwards this pruning is not required, which we claimed as an efficient process compared to other similar techniques like BIDE. In first iteration, we project an itemset s_p that spawned from selected itemset s_i from DB_S and an element e_i considered from 'I'. If the $f_{ts}(s_p)$ is greater or equal to rs, then an edge will be defined between s_i and e_i. If $f_{ts}(s_i) \cong f_{ts}(s_p)$ then we prune s_i from DB_S. This pruning process required and limited to first iteration only.

From second iteration onwards project the itemset S_p that spawned from $S_{p'}$ to each element e_i of 'I'. An edge can be defined between $S_{p'}$ and e_i if $f_{ts}(s_p)$ is greater or equal to rs. In this description $S_{p'}$ is a projected itemset in previous iteration and eligible as a sequence. Then apply the fallowing validation to find closed sequence.

If any of $f_{ts}(s_{p'}) \cong f_{ts}(s_p)$ that edge will be pruned and all disjoint graphs except s_p will be considered as closed sequence and moves it into DB_S and remove all disjoint graphs from memory.

If $f_{ts}(s_{p'}) \cong f_{ts}(s_p)$ and there after no projection spawned then s_p will be considered as closed sequence and moves it into DB_S and remove $s_{p'}$ *and* s_p from memory.

The above process continues till the elements available in memory those are connected through direct or transitive edges and projecting itemsets i.e., till graph become empty

1) Algorithm used in CEG&REP:

This section describes algorithms for initializing sequence database with single elements sequences, spawning itemset projections and pruning edges from Sequence Graph SG.

Algorithm 1: Concurrent Edge Prevision to build graph structure and Rear Edge Pruning

Step 1:

Input: Set of Elements 'I'.

Begin:

L1: For each element e_i of 'I'

Begin:

Find $f_{ts}(e_i)$

If $f_{ts}(e_i) \geq rs$ then

Move e_i as sequence with single element to DB_S

End: L1.

End.

Step 2:

Input: DB_S and 'I';

L1: For each sequence s_i in DB_S

Begin:

L2: For each element e_i of 'I'

Begin:

C1: if $\text{edgeWeight}(s_i, e_i) \geq rs$

Begin:

Create projected itemset s_p from (s_i, e_i)

If $f_{ts}(s_i) \cong f_{ts}(s_p)$ then prune s_i from DB_S

End: C1.

End: L2.

End: L1.

L3: For each projected Itemset s_p in memory

Begin:

$s_{p'} = s_p$

L4: For each e_i of 'I'

Begin:

Project s_p from $(s_{p'}, e_i)$

C2: If $f_{ts}(s_p) \geq rs$

Begin

Spawn SG by adding edge between $s_{p'}$ *and* e_i

End: C2

End: L4

C3: If $s_{p'}$ not spawned and no new projections added for $s_{p'}$

Begin:

Remove all duplicate edges for each edge weight from $s_{p'}$ and keep edges unique by not deleting most recent edges for each edge weight.

Select elements from each disjoint graph as closed sequence and add it to DB_S and remove disjoint graphs from SG.

End C3

End: L3

If $SG \neq \phi$ go to L3

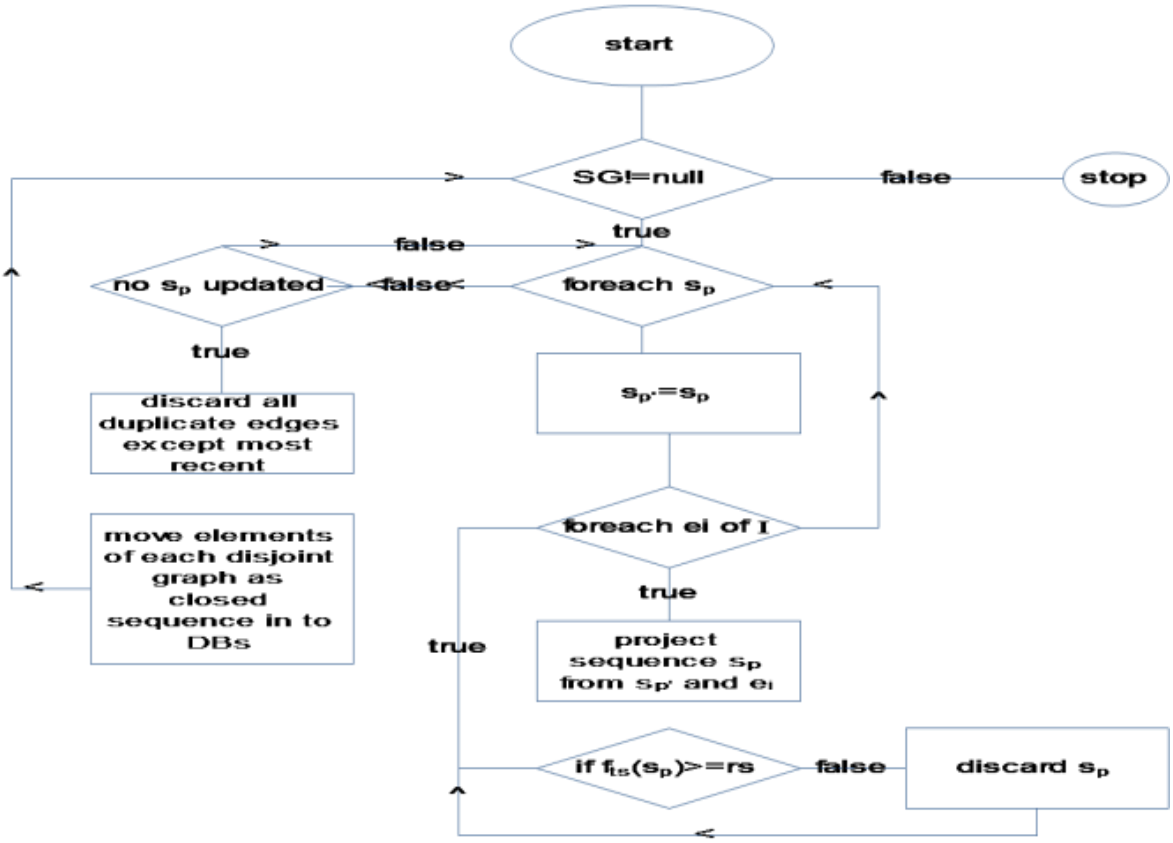


Fig 1: Concurrent Edge Prevision to build graph structure and Rear Edge Pruning

VI. COMPARATIVE STUDY:

This segment focuses mainly on providing evidence on asserting the claimed assumptions that 1) The CEG&REP is similar to BIDE which is actually a sealed series mining algorithm that is competent enough to momentously surpass results when evaluated against other algorithms such as CloSpan and spade. 2) Utilization of memory and momentum is rapid when compared to the ColSpan algorithm which is again analogous to BIDE. 3) There is the involvement of an enhanced occurrence and a probability reduction in the memory exploitation rate with the aid of the trait equivalent prognosis and also rim snipping of the CEG&REP. This is on the basis of the surveillance done which concludes that CEG&REP's implementation is far more noteworthy and important in contrast with the likes of BIDE, to be precise.

JAVA 1.6_ 20th build was employed for accomplishment of the CEG&REP and BIDE algorithms. A workstation equipped with core2duo processor, 2GB RAM and Windows XP installation was made use of for investigation of the algorithms. The parallel replica was deployed to attain the thread concept in JAVA.

A. *Dataset Characteristics:*

Pi is supposedly found to be a very opaque dataset, which assists in excavating enormous quantity of recurring clogged series with a profitably high threshold somewhere close to 90%. It also has a distinct element of being enclosed with 190 protein series and 21 divergent objects. Reviewing of serviceable legacy's consistency has been made use of by this dataset. Fig. 5 portrays an image depicting dataset series extent status.

In assessment with all the other regularly quoted forms like spade, prefixspan and CloSpan, BIDE has made its mark as a most preferable, superior and sealed example of mining copy, taking in view the detailed study of the factors mainly, memory consumption and runtime, judging with CEG&REP.

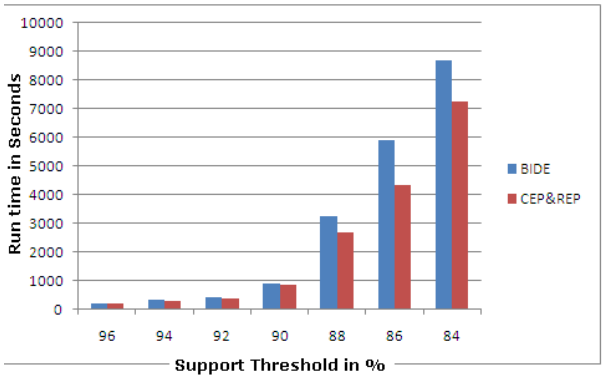


Fig 3: A comparison report for Runtime

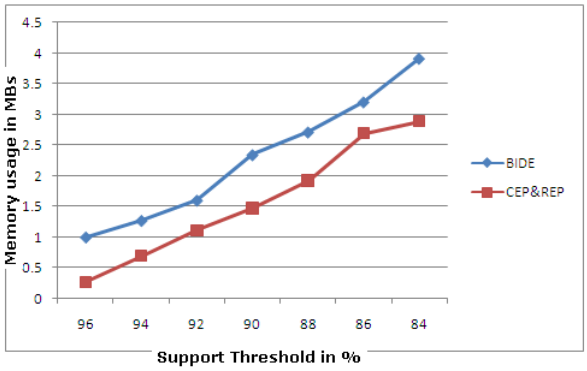


Fig: 4: A comparison report for memory usage

In contrast to CEG&REP and BIDE, a very intense dataset Pi is used which has petite recurrent closed series whose end to end distance is less than 10, even in the instance of high support amounting to around 90%. The diagrammatic representation displayed in Fig 3 explains that the above mentioned two algorithms execute in a similar fashion in case of support being 90% and above. But in situations when the support case is 88% and less, then the act of CEG&REP surpasses BIDE's routine.

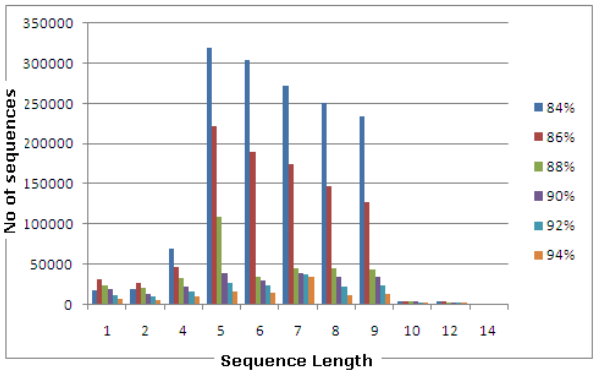


Fig 5: Sequence length and number of sequences at different thresholds in Pi dataset

The disparity in memory exploitation of CEG&REP and BIDE can be clearly observed because of the consumption level of CEG&REP being low than that of BIDE.

V. CONCLUSION

It has been scientifically and experimentally proved that clogged prototype mining propels dense product set and considerably enhanced competency as compared to recurrent prototype of mining even though both these types project similar animated power. The detailed study has verified that the case usually holds true when the count of recurrent moulds is considerably large and is the same with the recurrent bordered models as well. However, there is the downbeat in which the earlier formed clogged mining algorithms depend on chronological set of recurrent mining outlines. It is used to verify whether an innovative recurrent outline is blocked or else if it can nullify few previously mined blocked patterns. This leads to a situation where the memory utilization is considerably high but also leads to inadequacy of increasing seek out space for outline closure inspection. This paper anticipates an unusual algorithm for withdrawing recurring closed series with the help of Sequence Graph. It performs te following functions: It shuns the blight of contender's maintenance and test exemplar, supervises memory space expertly and ensures recurrent closure of clogging in a well-organized manner and at the same instant guzzling less amount of memory plot in comparison with the earlier developed mining algorithms. There is no necessity of preserving the already defined set of blocked recurrences, hence it very well balances the range of the count of frequent clogged models. A Sequence graph is embraced by CEG&REP and has the capability of harvesting the recurrent clogged pattern in an online approach. The efficacy of dataset drafts can be showcased by a wide-spread range of experimentation on a number of authentic datasets amassing varied allocation attributes. CEG&REP is rich in terms of velocity and memory spacing in comparison with the BIDE and CloSpan algorithms. ON the basis of the amount of progressions, linear scalability is provided. It has been proven and verified by many scientific research studies that limitations are crucial for a number of chronological outlined mining algorithms. Future studies include proposing of claiming a deduction advance on perking up the rule coherency on predictable itemsets.

REFERENCES

[1] F. Masseglia, F. Cathala, and P. Poncelet, The psp approach for mining sequential patterns. In PKDD'98, Nantes, France, Sept. 1995.

[2] R. Srikant, and R. Agrawal, Mining sequential patterns: Generalizations and performance improvements. In EDBT'96, Avignon, France, Mar. 1996.

[3] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.C. Hsu, FreeSpan: Frequent pattern-projected sequential pattern mining . In SIGKDD'00, Boston, MA, Aug. 2000.

[4] M. Zaki, SPADE: An Efficient Algorithm for Mining Frequent Sequences. Machine Learning, 42:31-60, Kluwer Academic Pulishers, 2001.

[5] J. Pei, J. Han, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.C. Hsu, PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In ICDE'01, Heidelberg, Germany, April 2001.

[6] J. Ayres, J. Gehrke, T. Yiu, and J. Flannick, Sequential PAttern Mining using a Bitmap Representation. In SIGKDD'02, Edmonton, Canada, July 2002.

[7] M. Garofalakis, R. Rastogi, and K. Shim, SPIRIT: Sequential PAttern Mining with regular expression constraints. In VLDB'99, San Francisco, CA, Sept. 1999.

[8] J. Pei, J. Han, and W. Wang, Constraint-based sequential pattern mining in large databases. In CIKM'02, McLean, VA, Nov. 2002.

[9] M. Seno, G. Karypis, SLPMiner: An algorithm for finding frequent sequential patterns using lengthdecreasing support constraint. In ICDM'02,, Maebashi, Japan, Dec. 2002.

[10] H. Mannila, H. Toivonen, and A.I. Verkamo, Discovering frequent episodes in sequences . In SIGKDD'95, Montreal, Canada, Aug. 1995.

[11] B. Ozden, S. Ramaswamy, and A. Silberschatz, Cyclic association rules. In ICDE'98, Olando, FL, Feb. 1998.

[12] C. Bettini, X. Wang, and S. Jajodia, Mining temporal relationals with multiple granularities in time sequences. Data Engineering Bulletin, 21(1):32-38, 1998.

[13] J. Han, G. Dong, and Y. Yin, Efficient mining of partial periodic patterns in time series database. In ICDE'99, Sydney, Australia, Mar. 1999.

[14] J. Yang, P.S. Yu, W. Wang and J. Han, Mining long sequential patterns in a noisy environment. In SIGMOD' 02, Madison, WI, June 2002.

[15] N. Pasquier, Y. Bastide, R. Taouil and L. Lakhal, Discoving frequent closed itemsets for association rules. In ICDT'99, Jerusalem, Israel, Jan. 1999.

[16] M. Zaki, and C. Hsiao, CHARM: An efficient algorithm for closed itemset mining. In SDM'02, Arlington, VA, April 2002.

[17] X. Yan, J. Han, and R. Afshar, CloSpan: Mining Closed Sequential Patterns in Large Databases. In SDM'03, San Francisco, CA, May 2003.

[18] J. Wang, J. Han, and J. Pei, CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets. In KDD'03, Washington, DC, Aug. 2003.

[19] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In VLDB'94, Santiago, Chile, Sept. 1994.

[20] J. Pei, J. Han, and R. Mao, CLOSET: An efficient algorithm for mining frequent closed itemsets . In DMKD'01 workshop, Dallas, TX, May 2001.

[21] J. Han, J. Wang, Y. Lu, and P. Tzvetkov, Mining Top- K Frequent Closed Patterns without Minimum Support. In ICDM'02, Maebashi, Japan, Dec. 2002.

[22] P. Aloy, E. Querol, F.X. Aviles and M.J.E. Sternberg, Automated Structure-based Prediction of Functional Sites in Proteins: Applications to Assessing the Validity of Inheriting Protein Function From Homology in Genome Annotation and to Protein Docking. Journal of Molecular Biology, 311, 2002.

[23] R. Agrawal, and R. Srikant, Mining sequential patterns. In ICDE'95, Taipei, Taiwan, Mar. 1995.

[24] I. Jonassen, J.F. Collins, and D.G. Higgins, Finding flexible patterns in unaligned protein sequences. Protein Science, 4(8), 1995.

[25] R. Kohavi, C. Brodley, B. Frasca, L.Mason, and Z. Zheng, KDD-cup 2000 organizers' report: Peeling the Onion. SIGKDD Explorations, 2, 2000.

[26] Jianyong Wang, Jiawei Han: BIDE: Efficient Mining of Frequent Closed Sequences. ICDE 2004: 79-90

[27] Anurag Choubey, Dr. Ravindra Patel and,Dr. J.L. Rana. Article: Frequent Pattern Mining With Closeness Considerations: Current State Of The Art. GJCST Issue 11, Volume 17, August 2011. Published by Global Journals, 25200 Carlos Bee Blvd. #495, Hayward, CA 94542, USA

AUTHORS PROFILE

Anurag Choubey[1]: He has completed B.Sc. (Electronics) in 1990 from Dr. Hari Singh Gaur Vishwavidyalaya, Sagar, M.P., India and M.Sc. (Applied Physics) in 1993 from Govt. Engineering College, Jabalpur, M.P., India. He has completed MCA in 2008 from Guru Ghasidas Vishwavidyalaya (Presently central University), Bilaspur C.G., India. He has worked as a Lecturer at Govt. Engg. College, Jabalpur and Hitkarni College of Engineering & Technology, Jabalpur from 1998 to 2000. From October 2000 onwards he has been working in Technocrats Institute of Technology, Bhopal. Currently working as Dean Academic, Technocrats Institute of Technology, Bhopal M.P., India. He posses more than 13 years of experience in teaching and has worked in different capacities like controller of exam, admission in-charge and other administrative post. email: anuragphd11@gmail.com

Dr. Ravindra Patel[2]: Associate Professor and Head, Department of Computer Applications at Rajiv Gandhi Technological University, Bhopal, India. He has awarded Ph.D. degree in Computer Science. He possessed more than 10 years of experience in teaching post-graduate classes. He has published more than 15 papers in international and national journals and conference proceedings. He is member of International Association of Computer Science and Information Technology (IACSIT). Email: ravindra@rgtu.net

Dr. J.L. Rana[3]: Former Professor and Head, Department of Computer Science and Engineering & Information Technology at MANIT, Bhopal, India with 32 years of vast experience of teaching, out of which 19 years as Professor of Computer Science & Engg. and Information Technology. Currently he is working as Group Director of Radha Raman Group of Institute, Bhopal. He has completed B.E. (Hons) in 1968 from GEC, Jabalpur and M.S. Computer Control from University of Hawaii (UH), USA in 1972. He has awarded Ph.D. degree in Computer from Indian Institute of Technology, Bombay in 1987. He also posses more than 25 years of experience in post graduate teaching He has published more than 30 papers in international and national journals and 65 international and national conference proceedings. He has guided 12 nos. of Ph.D. and 5 no. in progress He is a senior life member of CSI and Chairman CSI Bhopal Chapter for two terms. Email: jl_rana@yahoo.com

Error Filtering Schemes for Color Images in Visual Cryptography

Shiny Malar F.R
Dept. of Computer Science & Engineering
Noorul Islam University, Kumaracoil
Kanyakumari district, India

Jeya Kumar M.K
Professor, Dept. of Computer Applications
Noorul Islam University, Kumaracoil
Kanyakumari District, India

***Abstract* - The color visual cryptography methods are free from the limitations of randomness on color images. The two basic ideas used are error diffusion and pixel synchronization. Error diffusion is a simple method, in which the quantization error at each pixel level is filtered and fed as the input to the next pixel. In this way low frequency that is obtained between the input and output image is minimized which in turn give quality images. Degradation of colors are avoided with the help of pixel synchronization. The proposal of this work presents an efficient color image visual cryptic filtering scheme to improve the image quality on restored original image from visual cryptic shares. The proposed color image visual cryptic filtering scheme presents a deblurring effect on the non-uniform distribution of visual cryptic share pixels. After eliminating blurring effects on the pixels, Fourier transformation is applied to normalize the unevenly transformed share pixels on the original restored image. This in turn improves the quality of restored visual cryptographic image to its optimality. In addition the overlapping portions of the two or multiple visual cryptic shares are filtered out with homogeneity of pixel texture property on the restored original image. Experimentation are conducted with standard synthetic and real data set images, which shows better performance of proposed color image visual cryptic filtering scheme measured in terms of PSNR value (improved to 3 times) and share pixel error rate (reduced to nearly 11%) with existing grey visual cryptic filters. The results showed that the noise effects such as blurring on the restoration of original image are removed completely.**

Keywords - Error Diffusion; Visual Cryptography; Fourier Filtering; Context Overlapping; Color Extended Visual Cryptography.

I. INTRODUCTION

Visual Cryptography, an encryption technique allows cryptic to be possible only if the proper key is supplied by the user and decryption can be performed without the intervention of the computer. It works on the principle that when an image is splited into k shares only the user who has all the k shares can decrypt the message, any k-1 shares held by the user do not contain any useful information[1].

Naor and Shamir [2], in 1994 proposed a new security technique named visual cryptography scheme. In this technique, a secret image of type binary is encoded in a cryptographical manner into random binary patterns which contains n shares in a k-out-of-n scheme. The n shares are distributed among n participants in such a way the each participant's share is not known to another participant. The secret image can be visually revealed by k or more participants by joining all the shares available. Even if computational power decoding is available, cannot be done on the secret image by k-1 or fewer participants.

As the shares in the layers occur as random noise, the attackers cannot identify any useful information about the individual shares. Even with the availability of computer, it is not possible to decrypt the message or information with the limited availability of the share. The limitation of the above method is its randomness without any visual information. Extended Visual Cryptography have been suggested which also suffers from the same drawbacks of randomness. This paper is well thought-out as follows, Section II deals with the review of literature. Section III described about the error filtering schemes for color images. Section IV and V offered to Experimental result and discussion .Finally the conclusion of this paper in Section VI.

II. RELATED WORKS

Recently in the literature, many new methods have been implemented for visual cryptography. In 1995 Naor and Shamir [3], have predicted an optimal dissimilarity in k-out-of-n scheme to alleviate the contrast loss problem in the reconstructed image. A visual cryptography scheme is a broad spectrum method which is based upon general access structure. In k-out-of-n secret sharing scheme, any k shares will decode the secret image, which reduce the security level. To overcome this problem the basic secret sharing scheme is extended to general access structure. The concept of general access structure method was introduced in the year 1996 and 1997, by Ateniese, C.Blundo, A.Desantis and D.R.Stinson [4 , 5,6,7].In 1999,[8,9] Image size invariant visual cryptography was introduced by R. Ito, H. Kuwakado, and H. Tanaka and also in the same year the C.-N. Yang and C.-S. Laih have proposed some new types of visual secret sharing schemes.

In previous works of visual cryptography, binary images were concentrated which is not enough in real time applications. This general access structure method is applied to the gray level images are introduced by L. A. MacPherson,Chang Choulin[10,11,12],in the year 2000. In 2001 the G. Ateniese, C. Blundo, A. Santis, and D. R. Stinson have predicted the extended capabilities for visual cryptography in the natural images [13-16]. Ateniese has projected the hypergraph coloring method for Visual

cryptography , which is used to construct meaningful binary shares. Since hypergraph colorings are constructed by random distributed pixels, this method produce insufficient results. A new method of Extended visual cryptography for natural images is used to produce meaningful binary shares which is predicted by Nakajima[17,18] in the year 2002.Wen-Hsiang Tsai[19-23] have estimated the dithering technique which is applied to gray level images in visual cryptography. This technique is used for transformation of gray level images into binary images in the year 2003 . Again, Hou[24,25] has proposed the binary visual cryptography scheme which is applied to gray level images, that a gray level image is converted into halftone images in the year 2004.

In 2006 the Zhi Zhou, Gonzalo, R.Arce and Giovanni Dicrescenzo [29-33] have proposed halftone visual cryptography which produce good quality and meaningful halftone shares, the generated halftone shares contain the visual information. In halftone visual cryptography a secret binary pixel 'P' is encoded into an array of Q1 x Q2 ('m' in basic model) sub pixels, referred to as halftone cell in each of the 'n' shares. By using halftone cells with an appropriate size, visually pleasing halftone shares can be obtained and also maintained contrast and security. Abhishek parakh and Subhash Kak have proposed recursive threshold visual cryptography which is used in network applications and also reduce the network load. In 2007 the C.M. Hu and W.G. Tzeng [34, 35] have proposed a cheating method in Visual Cryptography schemes. In their cheating method, the cheater needs to know the exact distribution of black and white sub pixels of the shares of honest participants. In the same year, a Cheating Prevention Scheme for Binary Visual Cryptography with Homogeneous Secret Images was introduced by D.S. Tsai, T.H. Chen, G. Horng, Which is used to prevent the cheater from obtaining the distribution [26, 27, 28].

However, the knowledge of distribution is not a necessary condition for a successful cheats. They also proposed another cheat-preventing method in which the stacking of the genuine share and verification share reveals the verification image in some small region that it is possible to attack the method. Niranjan Damera-Venkata, and Brian L. Evans have predicted the design and analysis of vector color error diffusion halftoning systems. And also quantization of accumulate diffused errors in error diffusion method was introduced by Ti-Chiun Chang and Jan P. Allebach in the year 2005 [26, 27, 28].

In 2009 the Zhongmin Wang, Gonzalo R. Arce,, and Giovanni Di Crescenzo [36,37] have proposed the Visual Cryptography for color image using visual information pixel (VIP) synchronization with error diffusion technique. They are introduced a color Visual Cryptography encryption method which leads to significant shares and is free of the previously mentioned limitations. This method is used to filtering the error in an image and produces the meaningful shares. The error filtering schemes for color images is very simple and efficient method.

III. ERROR FILTERING SCHEMES FOR COLOR IMAES

A. Fourier filtering for color visual cryptographic images

The Fourier Transform of an image can be carried out using the Discrete Fourier Transform (DFT) method. Fig.1 shows the DFT also allows spectral data (i.e. a transformed image) to be inverse transformed, producing an image once again. If we compute the DFT of an image, then immediately inverse transform the result, we expect to regain the same image. If we multiply each element of the DFT of an image by a suitably chosen weighting function we can accentuate certain frequency components and attenuate others. The corresponding changes in the spatial form can be seen after the inverse DFT has been computed.

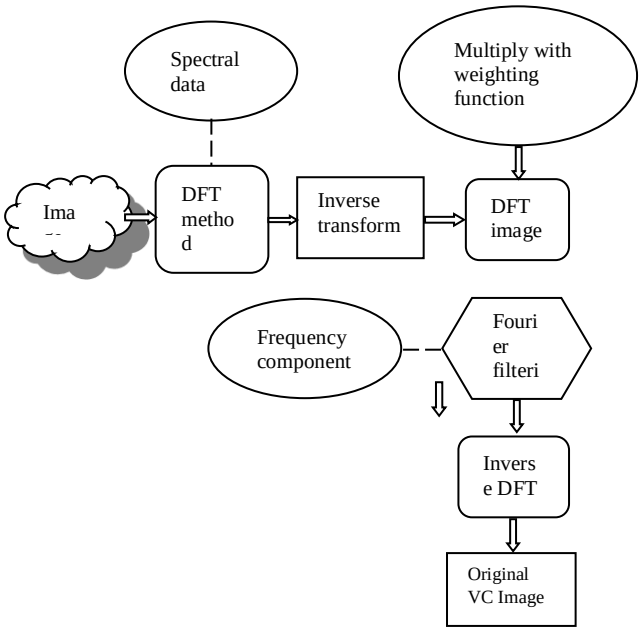


Figure 1. Fourier filtering for color visual cryptographic images.

The selective enhancement/suppression of frequency components is known as Fourier Filtering. The fourier filtering is used for convolution with large masks (Convolution Theorem), compensate for known image defects (restoration), reduction of image noise, suppression of 'hum' or other periodic interference and reconstruction of original restored visual cryptographic image.

1) Fourier Filtering

The DFT is the sampled Fourier Transform and does not have all frequencies to form an image, but only a set of forms which is large enough to fully define the spatial domain image. The total number of frequencies correspond to the total number of pixels in the spatial domain image, ***i.e.*** the image in the spatial and Fourier domain is of the equal size.

For a square image of size $N \times N$, the two-dimensional DFT is shown in the equation 1.

$$F(x,y) = \sum_{p=0}^{N-1} \sum_{q=o}^{N-1} f(p,q)\, e^{-y2\Pi (xp/N + yq/N)} \quad (1)$$

where f(p,q) is the image in the spatial domain . The exponential term is the basic function corresponding to each point F(x,y)in the Fourier space. The value of each point F(x,y) is calculated by multiplying the spatial image with the corresponding base function and adding the result.

Similarly, the Fourier image can be re-transformed to the spatial domain. The inverse Fourier transform is exposed in the equation 2.

$$f(i,j) = 1/N^2 \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} F(x,y)\, e^{y2\Pi (xi/N + yj/N)} \quad (2)$$

Here, $\frac{1}{N^2}$ normalization term in the inverse transformation. Sometimes, this normalization is carried out for the forward transformation instead of the inverse transformation. To access the result for the above equations, a double sum has to be obtained for each image point. However, the Fourier Transform is given by equation 3.

$$F(x,y) = 1/N \sum_{j=0}^{N-1} K(x,j)\, e^{-y2\Pi pj/N} \quad (3)$$

where

$$K(x,j) = 1/N \sum_{i=0}^{N-1} f(i,j)\, e^{-y2\Pi xi/N}$$

By using these two equations, initially the spatial domain image is transformed into an intermediate image using ***N*** one-dimensional Fourier Transforms. This intermediate image is then transformed into the final image, again use ***N*** one-dimensional Fourier Transforms. Expressing the two-dimensional Fourier Transform in terms of a series of ***2N*** one-dimensional transform reduces the number of needed computationsB. *Texture overlapping*

Texture overlapping filters decide which parts of the input image to be patched into the output texture. After finding a good patch offset between two inputs, the computer is the best patch seam (the seam yielding the highest possible MRF likelihood among all possible seams for that offset). The two overlapped visual cryptic shares images are copied to the output, cut by max-flow/min-cut algorithm and then stitched together along optimal seams to generate a new output that is shown in fig.2. When filtering an overlapped texture, we want the generated texture to be perceptually similar to the original image. In this approach, the concept of perceptual similarity has been formalized by a Markov Random Field (MRF). It brings an accurate estimation of perceptual effect according to human's vision.

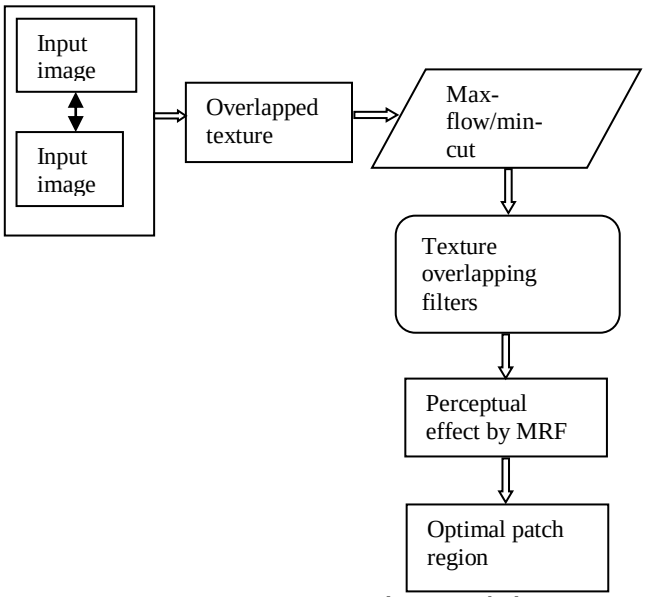


Figure 2. Texture overlapping method

In most of other techniques, the size of the patch is chosen a-prior. But this texture overlap filtering technique determine the optimal patch region for any given offset between the input and output texture. Finally the performance measure checks this flexibility for different offsets.

Let us assume a secret image *A* of N_R *X* N_M. Each pixel of *A* can take any one of M different colors or gray-levels. Image *A* is represented by an integer matrix A given by equation 4.

$$A = [a_{pq}]\ N_R \text{ X } N_M \quad (4)$$

Now *M =2* for a binary image, and *M =256* for a grayscale image with one byte per pixel. In a color image, the pixel value will be an index to a color table, thus *M = 256*. In a color image using an RGB model, each pixel has three integers: R (red), G (green) and B (blue). If each R, G or B takes value between 0 and 255, we have *M = 2563.*

The VCS requires taking pseudo-random numbers as input to guide the choice of the share matrices. Denote the share matrices in M_p *as* S_0^p *,....* $S^q_{|Mp|-1}$, and denote $P(S^p_q)$ *for p = 0, 1 and q = 0, 1,.....* $|Mp|-1$ as the probability that choosing the share matrix S^p_q .Hence the input of the pseudo-random numbers should guarantee ,that is represented as shown in the equation 5.

$$P(S^p_0) = P(S^p_1) = ... = P(S^p_{|Mp|-1}) \quad (5)$$

In order to choose a share matrix pseudo-randomly in *Mp* , the dealer needs at least $log_2 |Mp|$ bits pseudo-random numbers (we will consider the case that $log_2 |Mp|$ is not an integer in a later time). Denote *B(q)* as the binary representation of integer q with length $log_2 |Mp|$, i.e. *B(q)* is the binary string that represents q. Without loss of generality, we assume that when the input pseudo-random number is *B(q)*, the dealer chooses the share matrix S^p_q to encrypt the secret pixel p,and denote *P(B(q))* as the probability of generating the binary string B(q) by the pseudo-random generator. According to the equation 6,

$$P(B(0))=P(B(1))= ..... =P(B(|Mp|-1)) \quad (6)$$

In fact the cipher texts of the AES and Twofish have satisfied the above equation, because they have passed the serial test. Hence, take the AES and Twofish as the pseudo-random generator.

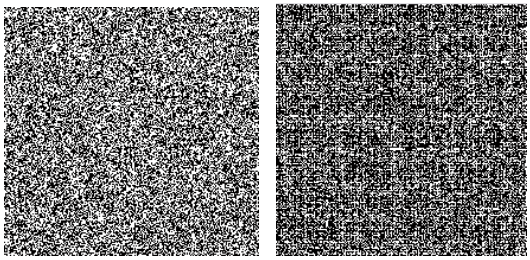
IV. EXPERIMENTAL RESULTS

In this paper, the experimental simulation is conducted by using the image processing software package (MATLAB). The color image (RGB image) is stored in MATLAB as an M-by- N-by-3 data array that defines red, green, and blue color components for every individual pixel. The color of each and every pixel is defined by the combination of the red, green, and blue intensities stored in each color plane at the pixel's location

a)Input image b) DFT image

Figure. 3 The experimental result of original input image with out Using error diffusion and DFT image using Fourier Filtering.

During the experiment, uncompressed image is taken as input image. Here used (2, 2) VCS scheme and consider the Lena color image of size 256 X 256 for experimental results shown in fig. 3(a). This input image is multiplied with the filter function in a pixel-by-pixel model. To have the resulting image in the spatial domain, filtered image has to be re-transformed using the inverse Fourier Transform. The most simple low pass filter is used to suppress all frequencies greater than the cut-off frequency and it leaves smaller frequencies unchanged. In most implementations, cut-off frequency is taken as a fraction of the highest frequency represented in the Fourier domain image shown in fig. 3(b).



(a)Share1 image (b) Share2 image

Figure. 4 Experimental result of (2,2) Visual cryptography Shares using error diffusion with the fourier Filtering method.

The (2, 2) VCS scheme is illustrated to introduce the basic concepts of texture overlapping schemes. In the encryption process every secret pixel is splitted into two shares. Each share belongs to the corresponding share image. In the decryption process the two corresponding shares are joined together by using OR operation to retrieve the secret pixel. Two share of a white secret pixel are of the equal while those of a black secret pixel are complementary as shown in Figure 4(a) and (b).

Figure 5. Decoded image from shares
Error diffusion with texture Overlapping

Consequently a white secret pixel is retrieved by a share with the combined result of half white sub-pixels and a black secret pixel is retrieved by all black. Using this basic VCS Scheme we can't completely retrieve the white Secrete pixel which generates loss in contrast. In XOR based VCS scheme where the share images are superimposed using XOR operation which results in perfect reconstruction of both Black and white pixels as shown in Figure 5 and sub sampling a 2 X 2 share into a single pixel we get decrypted image of the same size as original secret image.

The essential parameter indicates the superiority of the renovation is the Peak Signal-To-Noise Ratio(PSNR). PSNR is the ratio between the maximum possible power of the signal and the power of corrupted noise that is articulated in decibels.

Mean Square Error = Error/Size of the image (7)

The Mean Square Error is the average square of the error in particular images. The calculation of MSE & PSNR is given by the equation 8and equation 9.

$$MSE = 1/MN \left[\sum_{i=1}^{M} \sum_{j=1}^{N} (I_{ij} - I'_{ij})^2 \right] \quad (8)$$

$$PSNR = 20 * \log10 (255 / \text{sqrt}(MSE)) \quad (9)$$

Where, 255 is the maximum possible value of the image. In general the Peak signal -to-noise ratio for the two shares are increased and the perceived error for that two shares are decreased [38]. The imitation result also shows that the proposed scheme is compared to the existing scheme that is shown in table 1.

TABLE I. COMPARE THE EXISTING ERROR FILTERING METHOD AND PROPOSED ERROR FILTERING METHOD.

Error filtering Method	VC Scheme	Size of the Image	Test image - Lena	
			PSNR in dB	Error Ratio
Floyd & Steinberg Error Filtering (Existing)	2-out-Of-2	256 X 256	11.91	4.74
Discrete Fourier Filtering (Proposed)	2-out-Of-2	256 X 256	36.5826	0.0290

These works are some examples that prove the improvements and high performance of the color images in visual cryptography and also reduce the perceived errors.

V. RESULTS AND DISCUSSION

This section provides some experimental results to exemplify the effectiveness of the proposed method. The scheme proposed generates meaningful color shares with high quality as well as the colorful decrypted share by using Filtering scheme. The performance of the proposed method is evaluated and exposed in table.1 (that is, our proposed method is compared with the previous methods). VC can be treated as a special case in our proposed methods, which means no visual information is carried by the share. In existing method, shares carry visual information and there is a tradeoff between the contrast of the reconstructed image and the contrast of the share image. This tradeoff is similar to the tradeoff between the contrast of the reconstructed image and the image quality of the halftone shares in the proposed methods. Compared with the existing methods, our method achieves better image quality, which is given in table2.

Table 2. Reducing the error ratio of the images and Meaningful color shares with high visual Quality.

Existing Method				**Proposed Method**			
No.of Pixels	Error Rate	Color Shares	Error Rate	No. of Pixels	Error Rate	Color Shares	Error Rate
1000	11	500	22	1000	8	500	18
2000	19	750	29	2000	9	750	19
3000	28	1000	32	3000	13	1000	20
4000	38	1250	50	4000	19	1250	21
5000	48	1500	60	5000	23	1500	22

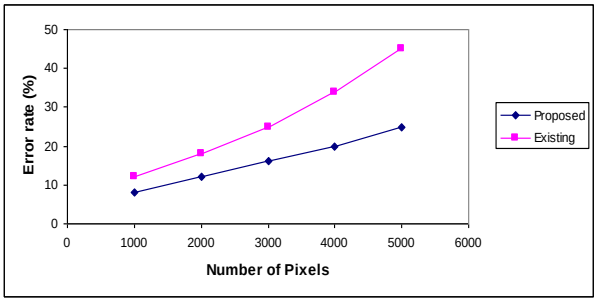


Figure 6. Number of pixels Vs Error rate

The results of experiments in which figure 6 and figure 7 indicate that the reducing the error ratio of the images and meaningful color shares with high visual quality that can improve the overall performance of the visual cryptography using texture overlapping and fourier filtering. The error rate is reduced to 11% compared with the existing scheme.

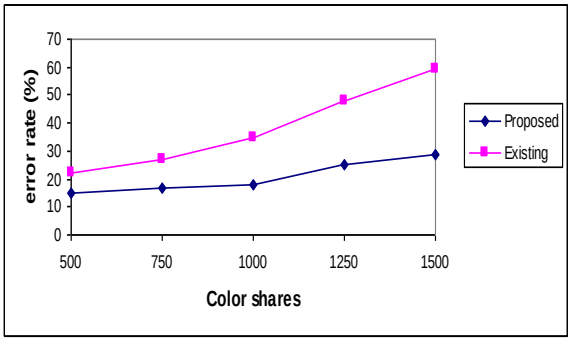


Fig 7. Color Shares Vs Error rate

VI. CONCLUSION

Some methods for color visual cryptography are not satisfactory in terms of producing either meaningless shares or meaningful shares with low visual quality, leading to suspicion of encryption. In the existing work of color VC the quality of images being restored depends on error diffusion, other image degradations due to blurring, transformation and overlapping were not handled in it.

The color VC focuses on the encryption method, to produce color Extended Visual Cryptographic System deploying VIP (Visual Information Pixel) Synchronization and Error Diffusion for improvement of quality. Error Diffusion results in the shares with good quality images and VIP Synchronization regains the actual values before and after encryption. This paper enhances the image quality on color visual cryptography using texture overlapping and Fourier filtering. The proposal in our work improves the image quality on restored original image from visual cryptic shares by presenting an efficient color image visual cryptic filtering scheme. The color image visual cryptic filtering method is presented here for deblurring effect on the non-uniform distribution of visual cryptic share pixels.

In the future, color image visual cryptic filtering scheme proposed in this paper, can be used to maintain digital document trade marking and licensing with ownership security schemes. Various multi-party security models used recently can be adapted in the future for the ownership security. Privacy preservation techniques (i.e., data transformation and perturbation) can also be considered for future direction in providing ownership confidentiality of digital documents.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their valuable advice and suggestions that have contributed to the improvement in the quality and presentation of this paper. In particular, we thank the reviewer who pointed us to error filtering method in this paper.

REFERENCES

[1] InKoo Kang, Member, IEEE, Gonzalo R. Arce, Fellow, IEEE, and Heung-Kyu Lee, Member, IEEE," Color Extended Visual Cryptography Using Error Diffusion," IEEE Transactions on image processing, vol. 20, no. 1, pp. 132-145, January 2011.

[2] M.Naor and A.Shamir, "Visual Cryptography ", in pro. EUROCRYPT, 1994,pp. 1-12.

[3] M.Naor and A. Shamir, Visual Cryptography, in "Advanced in Cryptology – EUROCRYPT'94", A. De. Santis, Ed., Vol. 950 of Lecture Notes in Computer Science, Springer-Verlag, Berlin, PP. 1-12,1995.

[4] G. Ateniese, C. Blundo, A. D. Santis, and D. R. Stinson, "Visual cryptographyfor general access structures," Inf. Comput., vol. 129, no. 2, pp. 86–106, 1996.

[5] G. Ateniese, C. Blundo, A. D. Santis, and D. R. Stinson, Extended Schemes for Visual Cryptography, submitted to Discrete Mathematics, 1996.

[6] M. Naor and B. Pinkas, "Visual authentication and identification," in Proc. Advances in Cryptology, 1997, vol. 1294, LNCS, pp. 322–336.

[7] E. R. Verheul and H. C. A. Van Tilborg, "Constructions and properties of k out of n visual secret sharing schemes," Designs, Codes, Cryptog,,vol. 11, no. 2, pp. 179–196, 1997.

[8] R. Ito, H. Kuwakado, and H. Tanaka, "Image size invariant visual cryptography," IEICE Trans. Fund. Electron. Commun. Comput. Sci., vol.E82-A, no. 10, pp. 2172–2177, 1999.

[9] C.-N. Yang and C.-S. Laih, "Some new types of visual secret sharing schemes," in Proc. Nat. Computer Symp., 1999, vol. 3, pp. 260–268.

[10] L. A. MacPherson, "Gray level visual cryptography for general access structrue," M. Eng. thesis, Univ. Waterloo, Ontario, Canada, 2000.

[11] T. Hofmeister, M. Krause, and H. U. Simon, "Contrast-optimal k out of n secret sharing schemes in visual cryptography," heoreti. Comput.Sci., vol. 240, pp. 471–485, 2000.

[12] C. N. Yang and C. S. Laih, "New colored visual secret sharing schemes," Designs, Codes Crypt., vol. 20, no. 3, pp. 325–336, 2000.

[13] G. Ateniese, C. Blundo, A. Santis, and D. R. Stinson, "Extended capabilities for visual cryptography," ACM Theor. Comput. Sci., vol. 250, pp. 143–161, 2001.

[14] G. Ateniese, C. Blundo, A. De Santis, and D. R. Stinson, "Extended Schemes for Visual Cryptography". Theoretical Computer Science, No. 250, pp. 143-161, 2001.

[15] C. Blundo, A. De Bonis and A. De Santis, ' Improved Schemes for Visual Cryptography". Designs, Codes, and Cryptography, No. 24, pp. 255–278, 2001.

[16] Niranjan Damera-Venkata, and Brian L. Evans, "Design and Analysis of Vector Color Error Diffusion Halftoning Systems"IEEE transactions on image processing, vol. 10, no. 10, pp. 1552 - 1565 October 2001.

[17] M. Nakajima and Y. Yamaguchi, "Extended visual cryptography for natural images," J. WSCG, vol. 10, no. 2, 2002.

[18] W.-G. Tzeng and C.-M. Hu, "Anewapproach for visual cryptography," Designs, Codes, Cryptog., vol. 27, no. 3, pp. 207–227, 2002.

[19] C. C. Lin and W. H. Tsai, "Visual cryptography for gray-level images by dithering techniques," Pattern Recognit. Lett., vol. 24, pp. 349–358, 2003.

[20] C. Blundo, P. D'Arco, A. De Santis, and D. R. Stinson, "Contrast Optimal Threshold Visual Cryptography Schemes' , SIAM J. on Discrete Math. 16, pp. 224-261, 2003.

[21] Y. T. Hsu and L. W. Chang, "A new construction algorithm of visual crytography for gray level images," in Proc. IEEE Int. Symp. Circuits Syst., 2006, pp. 1430–1433.

[22] Y. C. Hou, "Visual cryptography for color images," Pattern Recognit., vol. 36, pp. 1619–1629, 2003.

[23] C. C. Lin and W. H. Tsai, "Visual cryptography for gray-level images" Pattern Recognit. Lett., vol. 25, pp. 349–358,2003.

[24] M. Krause and H.-U. Simon, "Determining the optimal contrast for secret sharing schemes in visual cryptography," Combin., Probab. Comput., vol. 12, no. 3, pp. 285–299, 2003.

[25] C.-N. Yang, "New visual secret sharing schemes using probabilistic method," Pattern Recognit. Lett., vol. 25, no. 4, pp. 481–494, 2004.

[26] D. Jin, W. Q. Yan, and M. S. Kankanhalli, "Progressive color visual cryptography," J. Electron. Imag., vol. 14, no. 3, p. 033019, 2005.

[27] Ti-Chiun Chang and Jan P. Allebach, "Quantization of Accumulated Diffused Errors in Error Diffusion", IEEE Transactions on image processing, vol. 14, no. 12, pp. 1960 – 1975 , December 2005.

[28] C. N. Yang and T. S. Chen, "Aspect ratio invariant visual secret sharing schemes with minimum pixel expansion," Pattern Recogniti. Lett., vol.26, pp. 193–206, 2005.

[29] W. P. Fang and J. C. Lin, "Progressive viewing and sharing of sensitive images," Pattern Recogniti. Image Anal., vol. 16, no. 4, pp. 632–636,2006.

[30] Z. Zhou, G. R. Arce, and G. D. Crescenzo, "Halftone visual cryptography,"IEEE Trans. Image Process., vol. 18, no. 8, pp. 2441–2453, Aug. 2006.

[31] E. Myodo, S. Sakazawa, and Y. Takishima, "Visual cryptography based on void-and-cluster halftoning technique," in Proc. IEEE Int. Conf. Image Process., 2006, pp. 97–100.

[32] G. Horng, T.H. Chen, D.S. Tsai, "Cheating in Visual Cryptography," Designs, Codes and Cryptography, Vol. 38, No.2, pp. 219-236, 2006.

[33] S. J. Shyu, "Efficient visual secret image sharing for color images," Pattern Recognit., vol. 39, no. 5, pp. 866–880, 2006.

[34] C.M. Hu and W.G. Tzeng, "Cheating Prevention in Visual Cryptography," IEEE Transactions on Image Processing,Vol. 16, No. 1, pp. 36-45, 2007.

[35] D.S. Tsai, T.H. Chen, G. Horng, "A Cheating Prevention Scheme for Binary Visual Cryptography with Homogeneous Secret Images," Pattern Recognition, Vol. 40, No. 8, pp. 2356-2366, 2007.

[36] Zhen He ,'Hierarchical Error Diffusion", IEEE Transactions on image processing, Vol. 18, No. 7, pp. 1524-1534, July 2009

[37] Zhongmin Wang, Gonzalo R. Arce,, and Giovanni Di Crescenzo ,"Halftone Visual Cryptography Via Error Diffusion",IEEE Transactions on information forensics and security, Vol. 4, No. 3, 383-395, September 2009.

[38] Z. M. Wang, G. R. Arce, and G. Di Crescenzo, "Halftone visual cryptography via error diffusion," IEEE Trans. Inf. Forensics Security, vol. 4, no. 3, pp. 383–396, Sep. 2009.

AUTHORS PROFILE

F.R. Shiny malar was born in Nagercoil,Tamil Nadu State,India in 1986. She studied Information Technology in theSt.Xavier's Catholic college of Engineering, Chunkankadai, Kanyakumari District, Tamilnadu State,India fom 2003 to 2007. She received Bachelor‟ sdegree from Anna University, chennai 2007. And received the Master degree from Manonmaniam Sundaranar University Tirunelveli. currently, she is a research scholar at the Department of Computer Science and Engineering, in Noorul Islam Center for Higher Education, Noorul Islam University, Kumarakoil, Tamilnadu, India; working in the area of image processing under the supervision of Dr. M. K. Jeya Kumar. She has presented a number of papers in national conferences and their research interest include image security , networking and image processing.

M. K. Jeya Kumar received his PhD degree in Mobile Adhoc Networks from Dr. MGR University,Chennai, India, in 2010. He is Assistant Professor at the Department of Computer Application, Noorul Islam University, Kanyakumari District, Tamilnadu,India. His research interests include networks andnetwork security, image processing and softcomputing techniques.

Passwords Selected by Hospital Employees: An Investigative Study

B. Dawn Medlin
Computer Information Systems
Appalachian State University
Boone, NC, USA

Ken Corley
Computer Information Systems
Appalachian State University
Boone, NC, USA

B. Adriana Romaniello
Economía de la Empresa
Universidad Rey Juan Carlos
Madrid, Spain

***Abstract*— The health care industry has benefitted from its employees' ability to view patient data, but at the same time this access allows for patient's health care records and information to be easily tampered with or stolen. Access to and transmission of patient data may improve care, increase delivery time of services and reduce health care costs, security of that information may be jeopardized due to the innocent sharing of personal and non-personal data with the wrong person. In this study, we surveyed employees of different size hospitals in various regions of the state who were willing to share their passwords. Our findings indicate that employees need further or additional training in their awareness surrounding password creation.**

Keywords- Passwords; Security; HIPPA; HITECH.

I. INTRODUCTION

Health care records generally include, but are not necessarily limited to, individual patient's health history, diagnosis, laboratory results, treatments, and the doctor's progress notes. A patient's personal information, such as address, phone number, and social security number, are all items that may be included and accessible to some or all health care employees. These records are vulnerable to security breaches and theft. Both hackers and social engineers have successfully found ways to penetrate networked health data systems by simply asking for the information or by finding weaknesses within the system.

Unfortunately, the largest threat to a health care agency's security may not be outsiders, but rather their own employees. Inside employees actually can pose the largest threat to the security and privacy of information as they can exploit the trust of their co-workers, and they generally are the individuals who have or have had authorized access to the organization's network and who are familiar with its internal policies, procedures, and technologies. Additionally, internal employees can exploit that knowledge to facilitate attacks and even collude with external attackers (Insider Threat Research). Due to increased regulations and the increased opportunities for exploitation that exist in today's digital world, it is even more important for health care providers to keep health care records and the information held within, safe and private. Governmental agencies have adopted initiatives that specifically address the issues and rights of health care patients. More specifically, the security and privacy of health care information is protected by HITECH (Health Information Technology and Clinical Health Act) and the Health Insurance Portability and Accountability Act (HIPAA), requiring health care agencies to do everything possible to protect their information.

II. BACKGROUND

The electronic accumulation and exchange of personal health information has been promoted as a significant benefit to health care consumers and providers. Many health care policy experts believe that broader health information technology adoption may lead to the availability of more complete and transparent information, ultimately helping to contain health care costs while simultaneously improving health care quality.

Managers must be vigilant in their efforts to protect patient information as required by several laws. On February 17th, 2009, President Obama signed into law the Health Information Technology and Clinical Health Act (HITECH) as part of the American Recovery and Reinvestment Act. The HITECH Act enhances the security and privacy provisions as well as the penalties contained in the Health Insurance Portability and Accountability Act of 1996 (The Health Information Technology for

Economic and Clinical Health Act (HITECH Act): implications for the adoption of health information technology, HIPAA, and privacy and security issues, 2009). This new law also requires patients be notified in the event of a security breach.

In addition to HITECH, the basic goal of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) is to protect the privacy and security of patients and their medical records. Furthermore, HIPAA addresses security and privacy measures in relation to passwords, either directly or indirectly, in the following standards: 1) **Security Management Process [161.308(a)(1)]** Health care organizations must show that they have a consistent set of internal processes, with implementation that is widespread and institutionalized. Processes range from establishing criteria for who has access to what, and who can request certain resources; to ensuring that access rights are revoked immediately upon employee termination, 2) **Security Awareness and Training [161.308(a)(5)]** HIPAA requires that staff members be trained and educated concerning the

proper handling of PHI. This basic-level security training should include measures such as password management, and 3) **Access Control [161.312(a)]** HIPAA security regulations require a definition of who has access to PHI within the organization, as well as the rules determining an individual's right of access, and the reasons for denying access to some individuals.

Despite its legal requirements, however, HIPAA standards are not always followed. As an example, a public posting of 20,000 emergency room patients who had visited Stanford Hospital in Palo Alto, California, was placed on a commercial Web site that included the patient's names and diagnosis codes. The hospital confirmed that the information remained online for nearly a year (Sack, 2011). Another example included a laptop that was stolen from a rehabilitation center containing 660 patient's records. The laptop which was reported stolen from Rancho Los Amigos National Rehabilitation Center on Feb. 24, 2011 contained at least 667 patient names, their date of birth and diagnostic information (Downey, 2011). These are only two examples of some of the thousands of medical records that are either stolen or lost each year.

The Federal Trade Commission (FTC) and the Department of Health and Human Services (HSS) in 2009 issued the first set of HIPAA privacy/security guidance under the new HITECH Act requirements. The new guidance relates to the security breach notification requirement, that states "Under this requirement, health plans and personal health record (PHR) vendors must provide individual notification if there has been a security breach of protected health information" (http://compliance.utimaco.com/na/tag/hitech- act/).

Additionally, notification must be provided to individuals in writing within 60 days of discovery of the breach. If the breach involves more than 500 individuals, notice also must be made in prominent media outlets and to the Secretary of Housing and Health Services or to the FTC for PHR vendors (Health IT Data Breaches: No Harm, No Foul).

For health care administrators, security is enhanced by using systems tools that are already available, such as Active Directory and LDAP **(**Lightweight Directory Access Protocol**).** Most likely, one or the other, or a combination of both is already in use to help in the securing of information. Even when other front-end access management products, like IBM Tivoli, Citrix or Sun Microsystems' Java System Identity Manager are in use, the directory server on the back end is likely to be Active Directory, LDAP or both.

In addition, more health care agencies may consider adopting biometrics. Biometrics is the science of identifying people through physical characteristics. Usually not one technology but a cluster of several, biometrics uses fingerprints, handprints, retina scans, voice recognition, facial structure, and even hand motions while writing a signature-to identify individuals (Simpson, 2002).

Smart cards may also be used as these operate with a chip that includes stored memory, and an operating system. A patient's entire clinical history is stored on the smart card which can only be accessed via reading devices in a physician's office, primary care center, hospital, or other medical institution. Through the use of this device, exposed paper records will not be a concern. An added benefit of smart cards is the ability for users to electronically forward patient information to other health care authorities and insurers. Specifically, Java-based card technology emerges as a leading platform because of its ability to support multiple health care applications securely, while incorporating biometrics for positive identification and authentication.

A. Issues

Americans hold a strong belief in their right to privacy, and that belief has been served by the legal system of the United States. Privacy is also a constitutional concept, as found in the Fourth Amendment to the U.S. Constitution (Gostin, 2000). In fact, the preamble to the federal Privacy Rule, promulgated pursuant to HIPAA, notes that the existence of a generalized right to privacy as a matter of constitutional law suggests there are enduing values in American law related to privacy.

As required by HIPAA as well as other state laws, health care institutions are required to provide security methods in order to protect patient's information. One such method is through the authentication of the individual requesting access. Health care employees are generally subjected to some type of authentication process. Although there are different ways of authenticating employees, most systems are based on the use of a physical token (something one has), secret knowledge (something one knows) or biometrics (something one is) (Burnett & Kleiman, 2006).

In today's health care institutions, the most common authentication mechanism is still the simple use of a password (something one knows or creates). This type of authentication method can offer to employees the ability to quickly enter into a system, but human practices such as using the same password on different systems and writing down a password may degrade the quality of password security (Pfleeger and Pfleeger, 2007).

The authentication method of individuals creating their own passwords is not atypical. For health care organizations the password functions like the key to a lock, anyone who has it can get in to see the patient's information. Toward that end, there have been recommendations from governmental agencies to hospitals on how to construct a password. One of the first guidelines in creating good passwords was published in 1985 by the Department of Defense and is still relevant today (Department of Defense. 1985). Their guidelines recommended the following: 1) passwords must be memorized; 2) passwords must be at least six characters long, 3) passwords must be replaced periodically, and 4) passwords must contain a mixture of letters (both upper- and lowercase), numbers, and punctuation characters.

Most networks administrators and security experts would concur with all of the above Department of Defense recommendations, however, that was in 1985 when the advice was given and when social engineering as well as other types

of attacks were not as common as they are today. According to CERT (the Computer Emergency Response Team), the advice to use upper and lower case alpha characters for Novell and/or VMS systems is useless since both of these systems are case insensitive.

B. Problems

Many of the deficiencies of password authentication systems arise from the limitations of human cognitive ability (Pond et al., 2000). If humans were not required to remember a password, a maximally secure password would be one with maximum length that could consist of a string of numbers, character, and symbols. In fact, the requirements to remember long and complicated passwords are contrary to the way the human memory functions. First, the capacity of human memory in its capacity to remember a sequence of items is temporally limited, with a short-term capacity of around seven items plus or minus two (Kanaley, R., 2001). Second, when humans remember a sequence of items, those items cannot be drawn from an arbitrary and unfamiliar rang, but must be familiar 'chunks' such as words familiar symbols. Third, the human memory thrives on redundancy.

In fact, studies have shown that individuals' short term memory will retain a password for approximately 30 seconds thereby requiring individuals to attempt to immediately memorize their passwords. It has also been shown that if an individual is interrupted before they fully memorize the password; it will fall out of their working memory and most likely be lost.

Also, if an individual is in a hurry when the system demands a new password, individuals must sacrifice either the concentration of the critical task at hand or the recollection of the new password. Related to this issue is having to create the content for this new quickly demanded password. The pressure to choose creative and secure passwords quickly generally results in individuals failing in their attempt to memorize this new password. For health care organizations this can result in reset rates at one per reset per every four to five users per month (Brostoff and Sasse, 2001).

In order to combat the issue of having to remember so many different passwords some users have resorted to the selecting familiar terms such as a pet or family name, their own name, their phone number, or other common terms that could be found in a dictionary. British psychologist Helen Petrie, Ph.D., a professor of human/computer interaction at City University in London analyzed the passwords of 1,200 British office workers who participated in a survey funded by CentralNic, an Internet domain-name company in 2001. She found that most individuals' passwords fell into one of four distinct password categories which were family, fan, fantasists, and cryptic.

The first category of "family," comprised nearly half of the respondents. These individuals selected their own name, the name of a child, partner or pet, birth date, or significant number such as a social security number. Further, Dr. Petrie found that individuals also choose passwords that symbolized people or events with emotional value or ties.

One third of the survey participants were identified as "fans," using the names of athletes, singers, movie stars, fictional characters, or sports teams. Dr. Petrie also found that these individuals wanted to align themselves with the lifestyle represented by or surrounded around a celebrity status. Two of the most popular names were Madonna and Homer Simpson.

Fantasists made up eleven percent of survey responses. Dr. Petrie found that their passwords were comprised of sexual terms or topics. Some examples included in this category were terms such as "sexy," "stud" and "goddess."

The final ten percent of participants were identified as "cryptics." These users were seemingly the most security-conscious, but it should also be noted that they were also the smallest of all of the four identified categories. These individuals selected unintelligible passwords that included a random string of letters, numerals, and symbols such as Jxa+157.

Self-created computer passwords are generally personal, and they reflect the personalities of millions of people as they attempt to summarize their life through a few taps on the keyboard. As psychologists know, people and personalities are often very predictable in the aggregate (Andrews, 2004), as may be their choices of passwords. Psychologists have found that humans can store only five to nine random bits of information in their short-term memory, making it difficult to remember long and complicated passwords. Therefore, users have often chosen passwords with personal meanings that they can associate with something in their long-term memory.

III. Research Methodology

A. Data Collection

To obtain a fair statistical representation of the password security used in relation to health care organizations, a survey was given to employees at five hospitals of various sizes and in different regions of the state. Hospital administration approval was obtained, but the administration did not endorse the survey to respondents, nor did they ask them to participate. The data set was comprised of 118 responses. Data was gathered to not only determine how many employees would disclose their passwords and other personal information such as their address, phone number and email, but also simulated the types of information individuals were willing to share with co-workers, colleagues, or friends of colleagues. The information that employees were willing to share, including their passwords and other personal information, would certainly make it easier to hack into a system instead of having to "guess" at the necessary authentication information.

B. Analysis and Results

As seen in Table 1 (labeled password categories), half of the respondents created passwords consisting of family names, including their own name or nickname, the name of a child, or significant other. Interestingly, the findings noted in Table 2 indicate that most respondents were often required to use a password to access systems, but

rarely changed their passwords. As further indicated, most of the respondents used the same password on multiple accounts. The practice of rarely changing passwords and/or using the same password for multiple accounts would assist social engineers, thus allowing them to easily attain access to one system and possibly more.

TABLE 1. PASSWORD CATEGORIES

Variable Name	Question	Answers	N	Mean	Std Dev
Family	Does your password fit into this category?	1 = Yes, 0 = No	118	0.50	0.50
Cryptic	Does your password fit into this category?	1 = Yes, 0 = No	118	0.05	0.22
Number	Does your password fit into this category?	1 = Yes, 0 = No	118	0.45	0.50
Fan	Does your password fit into this category?	1 = Yes, 0 = No	118	0.15	0.95
Faith	Does your password fit into this category?	1 = Yes, 0 = No	118	0.03	0.18
School	Does your password fit into this category?	1 = Yes, 0 = No	118	0.02	0.13
Fantasy	Does your password fit into this category?	1 = Yes, 0 = No	118	0.00	0.00
Place	Does your password fit into this category?	1 = Yes, 0 = No	118	0.14	0.34
Other	Does your password fit into this category?	1 = Yes, 0 = No	118	0.51	0.50

Additionally, as seen above in Table 1, the largest category of password choices included some type of relationship to a family name being reported at fifty percent (50%). Fifteen percent (15%) of the respondents self-reported the inclusion of "fan-based" words, which could include names of athletes, singers, movie stars, and fictional characters or sports teams. "Place" was the next highest category, with fourteen percent (14%), using another identifiable piece of information such as the city where the employee works/lives.

The smallest of all of the self-identified password categories was "fantasy," followed closely by the categories of school and faith. Five percent (5%) of the employees selected the "cryptic" category, suggesting that these employees are security-conscious since that category includes passwords that are unintelligible.

TABLE 2. PASSWORD STATISTICS

Variable Name	Question	Answers	N	Mean	Std Dev
Pass_Freq	How often do use a password to access systems?	1= Very Often 5= Never	118	1.23	0.59
Pass_ Change	How often do you change your passwords?	1= Very Often 5= Never	117	2.85	1.13
Reuse	Most people use the same password on multiple accounts. How often do you do this?	1= Very Often 5= Never	118	2.47	1.32

As noted in Table 3, in the area of password and security training, most of the respondents, fifty-four percent (54%) indicated that their employer had offered password security training, with fifty-eight percent (58%) of the hospitals offering some other type of security awareness training (Table 3). Attendance by the employee in either a current password or security awareness training program was measured on a Likert scale of 1 being last week and 5 being never. The employees indicated that currently, they almost never attended the security awareness programs.

TABLE 3: PASSWORD TRAINING

VAR NAME	QUESTION	ANSWER	NO	MEAN	STD DEV
Pass_Train	Does your employer offer password security training?	1 = Yes, 0 = No	115	0.54	0.50
Awar_Train	Does your employer offer any other security awareness training?	1 = Yes, 0 = No	113	0.58	0.50
Current_Train	When was the last time you participated in either a password or another security awareness training program?	1= Last week, 5 = Never	115	4.09	1.08

IV. DISCUSSION

This study reveals several interesting findings. As noted earlier, most employees used the same passwords on multiple accounts, even though they frequently changed them. The actions of repeatedly using the same password are contrary to

suggested recommendations by most security experts, because a hacker who gained access to one account could more easily access other systems. Requiring individuals to maintain a new password for each system or application would obviously make systems more secure but is in conflict with humans' short-term human memory capabilities. Employees may consider it necessary to include familiar names, places, and numbers in their passwords so that they can easily recall them.

Though most employees indicated that their employers offered password security training either very often or often, it appears that either the types of training are not very effective or that the employees did not take it very seriously.

V. CONCLUSION

Findings of the present study indicate that employees are willing to share personal information with co-workers and friends of co-workers. Seventy-three percent (73%) of the employees shared information that a social engineer could use to create a profile of an employee and gain access to the employer's network and other confidential patient information. It is imperative that employees understand the consequences of sharing information as well as the importance of creating and maintaining strong passwords.

The simulation that was carried out during this study demonstrated that many employees may currently be in violation of HIPAA and HITECH regulations due to their willingness to share their information and their practice of creating weak passwords, thus allowing for easy access into a system. Hospitals and other health care agencies must identify ways to educate employees regarding HIPAA and HITECH regulations to protect patients and practices to create a long password, but on the other hand offers it freely to others. This study demonstrated that many employees may currently be in violation of HIPPA and HITECH regulations due to their willingness to create weak passwords and to share them with strangers through our survey instrument.

REFERENCES

[1] 6 Password Protection Tips That Every Computer User MUST Know!. Retrieved on October 23, 2011 from http://www.richtechgroup.com/business/2011/04/6-password-protection-tips-that-every-computer-user-must-know/

[2] Andrews, L.W. (2004). Passwords reveal your personality. Retrieved March 13, 2007, from http://cms.psychologytoday.com/articles/pto-20020101-000006.html.

[3] Brostoff, S., & Sasse, M. A. (2001). *Safe and Sound: a safety-critical approach to security*. Position paper presented at the New Security Paradigms Workshop 2001, Cloudcroft, New Mexico. [4] Burnett, M. & Kleiman, D. (2006). *Perfect Passwords. Selection, Protection, Authentication.* Syngress.

[5] Data Breach Harm Analysis from ID Analytics Uncovers New Patterns of Misuse Arising from Breaches of Identity Data. Retrieved on November 12, 2009, http://www.idanalytics.com/news_and_events/20071107.html

[6] Department of Defense. (1985). Password Management Guideline. http://www.alw.nih.gov/Security/FIRST/papers/password/dodpwman.txt

[7] Downey (2011). Laptop Stolen from Rehab Center with Over 660 Patient Records. KTLA News. Retrieved on October 23, 2011 from http://www.ktla.com/news/landing/ktla-laptop-stolen-from-downey-rehab,0,3270396.story

[8] Georgetown University Information Security. Retrieved November 12, 2009, from

[9] Gostin, L.O. (2000). Public Health Law: Power, Duty, Restraint. University of California Press , Berkeley, CA. pp. 132-134.

[10] Gragg, D. (2007). A Multi-Level Defense Against Social Engineering." SANS. http://www.sans.org/reading_room/whitepapers/engineering/920.php.

Human Memory. Intelegen, Inc. Retrieved on December 1, 2009 from http://brain.web-us.com/memory/human_memory.htm.

[11] Health Law Alert. Retrieved on November 15, 2009 from http://www.nixonpeabody.com/publications_detail3.asp?ID=2621.

[12] Health IT Data Breaches: No Harm, No Foul. Retrieved on Novemer 12, 2009 from http://compliance.utimaco.com/na/tag/hitech-act.

[13] Hupp, M. (2007). Protecting patient medical records from the nosy. Retrieved on November 30, 2009 from http://www.bizjournals.com/milwaukee/stories/2007/11/12/focus3.html?t=printable.

[14] Insider Threat Research. Retrieved December 1, 2009 from http://www.cert.org/insider_threat.

[15] Internet Identity Theft And Password Security Tips. Retrieved on October 23, 2011 from http://www.combat-identity-theft.com/internet-identity-theft.html

[16] Kanaley, R. (2001). Login error trouble keeping track of all your sign-ons? Here's a place to keep your electronic keys, but you better remember the password. *San Jose Mercury News*, 3G.

[17] Pfleeger, C.P. & Pfleeger, S.L. (2007). *Security in Computing*. Fourth edition. Prentice Hall.

[18] Pond, R., Podd, J., Bunnell, J., Henderson, R. (2000). "Word Association Computer Passwords: The Effect of Formulation Techniques on Recall and Guessing Rates," *Computers & Security*, 19, 645-656.

[19] PROTECTING HEALTH INFORMATION. RETRIEVED ON NOVEMBER 12, 2009 FROM HTTP://WWW.HHS.GOV/NEWS/FACTS/PRIVACY.HTML.

[20] Sack, K. (2011). Patient Data Posted Online in Major Breach of Privacy. New York Times. Retrieved on October 23, 2011 http://www.nytimes.com/2011/09/09/us/09breach.html?pagewanted=all)

[21] *Simpson*, R.L. (2002). Nursing Management. Chicago: 33(12), 46-48.

[22] The Health Information Technology for Economic and Clinical Health Act (2009). Retrieved on October 23, 2011 from http://www.nixonpeabody.com/publications_detail3.asp?ID=2621

[23] Thompson, S. T. Helping the Hacker? (2006) Library Information, Security, and Social Engineering. Information Technology and Libraries. Chicago: 25(4), 222-226.

AUTHORS PROFILE

B. Dawn Medlin is the Chair and Professor in the Department of Computer Information Systems and the Co-Director of the Center for Advanced Research on Emerging Technologies, John A. Walker College of Business, at Appalachian State University in Boone, NC. During her 24 years of teaching she has taught courses such as Web 2.0 Technologies in Business, Introduction to Gaming, Advanced Security, Issues in E-Commerce, She has published in journals such as The Journal of Information Systems Security, Information Systems Security, International Journal of Electronic Marketing and Retailing, and the International Journal of Healthcare Information Systems and Informatics. Additionally, she has taught at the Université d'Angers and Addis Ababa University in Ethiopia.

Ken Corley is an Assistant Professor of Computer Information Systems at Appalachian State University. He received his Ph.D. from Auburn University in Auburn, Alabama, USA. His current research interests include information privacy & security, computer and human interaction, and sustainability.

B. Adriana Romaniello, originally from Uruguay, is an Interim Associate Professor of Management at the Department of Business Administration at University Rey Juan Carlos (Madrid, Spain) who came from University Carlos III (Madrid) where she teaches Corporate finance. She holds a Ph. D. in Business Administration at University Complutense of Madrid and a Licentiate degree in economics and business from University of La Republica (Uruguay). She teaches management, organization theory and organizational design to undergrads and doctoral students. Adriana's research interests focus on information security, organization theory and competitive strategy.

Current Trends in Group Key Management

R. Siva Ranjani[1]
Research Scholar, Dept.of CS&SE
Andhra University, Visakhapatnam,
Andhra Pradesh, India,

Dr.D.Lalitha Bhaskari[2]
Associate Professor, Dept.of CS&SE
Andhra University, Visakhapatnam,
Andhra Pradesh, India,

Dr.P.S.Avadhani[3]
Professor, Dept.of CS&SE
Andhra University, Visakhapatnam,
Andhra Pradesh, India,

***Abstract*—Various network applications require sending data onto one or many members, maintaining security in the large groups is one of the major obstacles for controlling access. Unfortunately, IP multicast is not providing any security over the group communication. Group key management is a fundamental mechanism for secured multicast. This paper presents relevant group key management protocols. Then, we compared them against some pertinent performance criteria. Finally, we discuss the new research directions in group key management.**

Keywords-Multicast; group key management; security member driven; time driven.

I. INTRODUCTION

With rapid growth in the internet, people using the group communication in applications such as paying TV, transmission of video and audio, updating software, military applications, video games etc. In recent decades, the focus is mainly on the security issues involved in the group communication. When the group uses the unicast communication, one sender is sending the data stream onto one group member. In multicasting, the group member is sending the data onto other group members. Security is main focused area in group communication. Group key management is the fundamental mechanism provides the security in group communication. In this, security is achieved by sharing a common key among the group members. The message packets, those are going to transmit should be encrypted with the shared key.

Group key management is mainly focusing on the key generation and distribution of key among the group members. All the group members should participate in the secure distribution, creation and revocation of the keys [1]. The communication session in group key management is managed by two entities: Group Controller (GC), responsible for key generation, distribution and rekeying for membership change and Key Server (KS), responsible for maintaining the keys and distributing the keys.

The scenario of group communication is shown in the figure 1. Each member in the group is having two keys (TEK and KEK). The TEK (Traffic Encryption Key) is used for encrypt, decrypt and authenticate the data transfer. The TEK for the group member is generated by the local manager. The KEK (Key Encryption Key) is used for encrypt the TEK. To multicast the message (m) secretly the sender encrypts the message with TEK using the symmetric key algorithm. At the receiving side, the receiver decrypts the message (m) with TEK. In the group communication, the members in the group are not fixed, members can join / members can leave the group. So we need to secure the sending message to be received by the group members at that instance. When member is leaving, the KS must generate a new TEK and distribute the key secretly to all other members except leaving one. This process is known as rekeying. From the figure, we observe that Key Server is sharing a secret key called Key Encryption Key (KEKi) with each group member. When the member is leaving, the KS generates a new TEK : TEK1, encrypted with their KEKi and sends it to all other group members except leaving one. So the leaving member does not know the new TEK1, to decrypt the future messages shared in the group.

When a new member is joined in the group, first it must be authenticated by the GC. After that, the KS checks the rights of the new member and adds the member in the future message transformation session. The KS generates a new secret KEKj and shared with the new member mj. In order to restrict the new member form past data access the KS generate the new TEK : TEK1, encrypted with KEKi and then sends to all the group members along with the new joined one.

II. GROUP KEY MANAGEMENT PROTOCOL

As defined by Menezes et al. in [2], Group Key Management is the set of techniques and procedures used for the establishment and maintenance of keys among members to form the group. According to Hutchison [3], group key management can be classified into three categories.

Centralized Group Key Management Protocols—Key distribution is achieved by a single entity i.e Key Distribution Center (KDC), also known as Central Authority (CA). The Central Authority maintains the entire group, allocates the individual KEK to group members. It is also responsible for sharing the common TEK among all the group members.

Decentralized Group Key Management Protocols — In Decentralized Group, the group is splitting into several subgroups. Each subgroup is managed by subgroup controller. In this approach, the hierarchy of sub group controllers shares the labor in transferring TEK to group members. This management will reduce the load on the KDC.

Distributed Group Key Management Protocols — In Distributed Group key management either all the group members or only one member is involved in group key generation. No group controller is present; this will improves the reliability of the overall system.

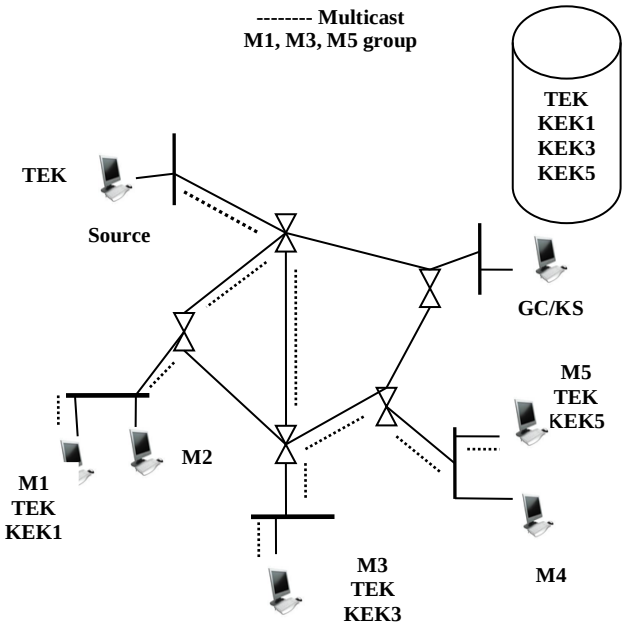


Figure 1: Group with KS and 5 Group Members

A. *Centralized GKMP*

Assume there is a group with n members. The Key server (KS) is the centralized group manager which stores information about all the group members. The KS takes n Key Encryption Keys (KEK) and each of them is shared with one member. The KEK is the secret key used for encrypting the group key (TEK). In the figure, the centralized group key protocols again sub divided into three categories depending on the technique used in distributing the TEK among group members.

a) Pair wise Keys:

In this sub type the KS shares a secret key called Key Encryption Key (KEK) with all the group members. This key is used for establishing the secured channel between the KS and the member for transferring the TEK securely whenever the key is required.

Harney and Muckenhirn [4],[5] proposed Group Key Management Protocol (GKMP), in this KS shares a secret key (KEK) individually with each active member and KS generates a Group Key Packet (GKP) that contains a Group TEK (GTEK) and a group KEK (GKEK). Chu et. al Protocol is proposed by Chu et. al [6], a Group leader shares a secret Key Encryption Key (KEK) with each group member.

b) Broadcast Secrets:

In this protocol, the KS broadcasts the rekey information to all the group members. Chiou and Chen [9] proposed Secure Lock protocol, in this key server uses a single broadcast to establish the group key or for sending rekey to the entire group in the case of join or leave membership.

c) Keys Hierarchy:

In pair wise key approach, the KS establishes individual secure channel with the members and uses this channel for sending the TEK updates. This mechanism increases the update message overhead. In order to reduce message overhead, the Key Server in this approach, shares a secret key with subgroup in addition to individual channel. When the member leaves the group the KS uses the sub group secret key to distribute the new TEK, which are not known by the leaved member. Nemaney et al.[26] proposed hierarchical group key management that increases the efficiency of the key. Following section describes some of the protocols using this re-key mechanism.

Wong et.al and Wallner et. al[11][14] proposed Logical Key Hierarchy (LKH) protocol. In this protocol, KS is the root of the tree and maintains a tree of keys. In this protocol, each node stores at most 1+log2(n) rekey messages.

McGraw and Sherman proposed One-Way Function Tree (OFT) protocol and this is an improvement over the LKH. Here, node's KEK is calculated by the member rather than attributed by the KS. Each node in this protocol is maintaining its blended sibling keys and its leaf secret key also maintains the blinded secret KEKs of its ancestors.

Canetti et. al proposed One-Way Function Chain Tree (OFCT) protocol [13]. This protocol works similar to OFT but, a pseudo random generator is used to generate the new KEKs rather than a one-way function, and it is done only during user removal.

Efficient Large Key distribution (ELK) [15] approach uses the pseudo random function for generating the new KEK when a membership change takes place. Waldvogel et.al [16] proposed Centralized Flat Table Key Management (CFKM) protocol, this approach, uses the flat table concept in order to reduce the number of keys maintained by the KS. Flat table consists of one TEK and 2w entries for KEKs, where w is the number of bits in identifier of a member. Wong et.al protocol is the extension of the LKH protocol [14]. The LKH uses the binary tree for key distribution; wong et.al uses the k-ary tree.

Comparison of centralized group key management protocols:

Table I compares the centralized group key management protocols. The efficiency of the protocol can be compared against the following criteria: 1 affect n, forward and backward secrecy, storage requirements at KS and group member, collusion, join re-key overhead and leave rekey overhead.

a) Decentralized Group key Management Protocols

The group members are arranged into some subgroups, and each subgroup has a controller called key manager. The key managers of the subgroup share the labor of distributing the TEK to group members in order to avoid bottle necks and single point of failure. Decentralized group key management is categorized into member ship driven and time driven re-keying.

Ballardie's Scalable Multicast Key Distribution (SMKD) protocol [18] propose a group key distribution method based on the Core Based Tree (CBT) multicast routing protocol. In CBT architecture, the multicast is rooted at main core. In Intra Domain Group Key Management (IGKMP) [17], the network divides into administratively scoped areas. This protocol is having Domain Key Distributor (DKD) and Area Key Distributor (AKD).

TABLE I : COMPARISON OF CENTRALIZED GROUP KEY MANAGEMENT

Protocol Type	Server Storage	Rekeying overhead	
		Member Join	Member leave
Poovendran et al	n+2	2	2n
Dunigan & chao	n+2	2	2n
Chu et. Al	n+2	2	2n
Secure Lock	2n	2	0
LKH	2n-1	$\log_2(n)+1$	$2\log_2(n)$
OFT	2n-1	$\log_2(n)+1$	$\log_2(n)+1$
OFCT	2n-1	$\log_2(n)+1$	$\log_2(n)+1$
ELK	2n-1	$\log_2(n)+1$	$\log_2(n)+1$
CFKM	2I+1	2I	2I

Where n: number of group members I: number of bits in a number id

The DKD is responsible for generating group TEK and is propagated to the group members through AKD. The DKD and AKDs belong to multicast group called All-KD-Group.

In Hydra protocol [19] the group is organized into sub groups. Each sub group has a server called Hydra Server (His) responsible for controlling the sub group. BAAL protocol has three entities: First is the Group controller (GC), responsible for maintaining the participant List (PL) and creating and sending the group key TEK to member through local controller. Second is Local Controller (LC), responsible for managing the keys in subnet, receives the new TEK and distributes to members connected to subnet. Third is Group member. IOLUS protocol is the frame work of a hierarchy of multicast subgroups to constitute virtual group [20]. Each subgroup is managed by a Group Security Agent (GSA), responsible for managing key inside the sub group.

Cipher Sequences is a proposed framework for multicast security [21], based on Reversible cipher sequence. The multicast tree is rooted at source and the leaves are group members. Challel et.al proposed Scalable Adaptive Key Management Scheme (SAKM) protocol. This protocol tackles the scalability issue. SAKM tackles the scalability by organizing the group into clusters.

b) Time Driven Approach

In time driven approach, the TEK is changed after specified amount of time. When the member leaves or joins in the group they will not excluded or appointed immediately, need to wait for the beginning of the new interval of time.

Briscoe proposed MARKS protocol, suggests a slicing the time length into small portions of time and uses a different key for encrypting each slice. The encryption keys occupied at leaves in BST that is generated from a single seed.

Setia et al [22] describe a scalable approach based on time-driven called Kronos. In this protocol, Setia denotes the group with a birth and death process model and discussed the model in two occasions: correlation subscriber behavior and independent subscriber. The operation of Kronos is similar to that of IGKMP. In Dual Encryption Protocol (DEP), the group is divided hierarchically into sub groups and the sub group is managed by sub-group manager (SGM). In YANG et. al Protocol [23] approach the multicast group is organized into a set of sub groups, KS manages each subgroup. The KS is responsible for rekeying the members in the subgroup periodically. Scalable Infrastructure For Multicast Key Management (SIM-KM) uses the proxy encryptions. SIM-KM uses the proxy function that converts the cipher text for one key into the cipher text for another key.

c) Comparison of Decentralized Group Key Management Protocols

In this different Group controllers are used to manage the subgroups. Table II compares the decentralized group key management protocols. Attributes that are used for evaluating the performance of decentralized protocols are key independence, decentralized controller, local rekey, key-data transformation, rekey per membership and type of communication.

TABLE II: COMPARISON OF DECENTRALIZED GROUP KEY MANAGEMENT

Protocol	**Key Independent**	**Decentralized Controller**		**Local re-key**	**Key Vs Data**	**Re-key**	**Communication Type**
		Management	**Key Server**				
SMKD	Yes	Yes	Yes	No	Yes	No	Both
IGKMP	Yes	Yes	Yes	No	Yes	Yes	Both
Hydra	Yes	Yes	Yes	No	Yes	Yes	Both
Baal	Yes	Yes	Yes	No	Yes	No	Both
MARKS	No	Yes	-	No	Yes	No	Both
Kronos	No	Yes	Yes	No	Yes	No	Both
DEP	Yes	No	No	No	Yes	No	Both
Iolus	Yes	Yes	Yes	Yes	Yes	Yes	1 to n
KHP	Yes	Yes	No	Yes	No	Yes	Both
Cipher Sequences	Yes	No	No	No	Yes	Yes	1 to n
SAKM	Yes	Yes	Yes	Yes	Yes	No	1 to n
Yang et al	Yes	Yes	Yes	Yes	Yes	Yes	1 to n
SIM-KM	No	No	No	Yes	Yes	Yes	1 to n

Both: 1 to n and n to n

d) Distributed Key agreement Protocols

In distributed key agreement protocol, the group members are participated in the establishment of a group key, further classified into three categories: Ring-based, hierarch based and broadcast based.

e) Ring Based

In this, cooperation of group members forms a virtual ring. In Ingemarson Et Al. protocol, all the group members are organized into a virtual ring and the Group Diffie-Hellman (GDH) protocol uses the extension of Diffie Hellman algorithm for group key generation.

f) Hierarchy based cooperation

The group members are arranged in a tree hierarchy for group key generation. In OCTOPUS, the entire group is divided into four sub groups. The leader member in the subgroup is responsible for collecting the intermediate subgroup values and calculates the intermediary DH value. Steer et. al proposed Skinny Tree (STR) protocol, uses the tree structure. The leaves associated in the tree are group members; each leaf is identified by its position. Diffie-Hellman Logical

Key Hierarchy (DH-LKH), proposed by Perrig et al.[24] is variant of STR and uses binary tree. The binary tree built from bottom to top. Distributed Logical Key Hierarchy (D-LKH) protocol uses the notion of sub-trees, agreeing on a mutual key. Distributed One-way Function Tree (D-OFT) approach using logical key hierarchy in a distributed fashion was proposed by Dondeti et al, uses the one-way function tree proposed by McGrew and Sherman. Every group member is trusted with access control and key generation. In Fiat and Naor protocol, each member broadcast a single message to other participants in order to agree on a common secret.

g) Broadcast based approach

In this approach, group key is generated by broadcasting the secret messages and distributing the computations among the group members. Burmester And Desmedt Protocol is a three round protocol with member generation, broadcasting and group key computations.

Boyd proposed Conference Key Agreement (CKA) protocol, where all the group members contributed to generate the group key.

h) Comparison of Distributed Group Key Management Protocols:

All the members in the group are involved in the computation of group key or generated by one member in the group. Table III compares the distributed group key management protocols. Attributes to evaluate the efficiency of distributed key management protocols are a number of rounds, number of messages, DH key and leader requirement.

TABLE III: COMPARISON OF DISTRIBUTED GROUP KEY MANAGEMENT

Protocol	No of Rounds	No of Messages		DH Key	Leader Req.
		Uni-cast	Multi-cast		
GDH	n	n-1	n	Yes	No
Ingemarson et. al	n-1	n(n-1)	0	Yes	No
Octopus	2(n-1)/4 – 2	3n-4	0	Yes	Yes
STR	n	0	N	Yes	No
DH-LKH	$\log_2(n)$	0	$\log_2(n)$	Yes	No
D-LKH	3	N	1	No	Yes
D-OFT	$\log_2(n)$	$2\log_2(n)$	0	No	No
D-CFKM	n	2n-1	0	No	No
Fiat et. Al	2	N	N	Yes	Yes
Bermester et	3	0	2n	No	No
CKA	3	n-1	N	No	Yes

III. CURRENT RESEARCH DIRECTIONS

A group Key Management application in mobile networks, ad hoc networks, e learning, and peer-to-peer networks is prevalent. Many new protocols are proposed as existing key management protocols are no more suitable for these areas. Jiang and Hu [8] classified current group key management protocols as stateless, self-healing, distributive, reliable, adaptive and mobile-based. Among these protocols reliability and distributiveness are by default provided by the group key management protocols. Scalability of stateless group key management protocols is enhanced by reducing the degree of the polynomial functions with the help of the decentralized subgroup managers. Junbeom and Hyunsoo [12] proposed a decentralized multi-group key management scheme for stateless group members. Self-healing and rekeying are becoming target areas in the group key management protocols. Key Server transmits group key updating messages when there are some changes in membership states. A self-healing protocol can recover certain number of existing and/or future group keys. First, self-healing key distribution protocol [8] was proposed by J.Staddon et.al [25] which is based on polynomials and Angelo [10] provided an efficient self-healing scheme for LKH. Challal et al.[7] proposed adaptive group key management protocol and there is need of extensive research should be done in this

IV. CONCLUSION

This paper focused on group key management, secured distribution of session keys and refreshment of the keying material. Reviewed so many group key management protocols and placing them into three main classes: centralized, decentralized and distributed protocols, which try to minimize the requirements of KDC and group members. Centralized key management is easy to implement but more overhead on single member. The decentralized key management follows the hierarchical sub grouping and it is harder to implement. Distributed key management is simply not scalable. From the comparison tables, we analyze that no unique solution that can achieve all the requirements. Hence, it is important to understand fully the requirements of the application before selecting a security solution. A solution for secure group communication should complement a multicast application rather than drive its implementation. The usage of security mechanism for secure group communication should be made transparent to the user and it should also work well with other protocols.

REFERENCES

[1] David Manz, Jim Alves-Foss and Shanyu Zheng, "Network Simulation of Group Key Management Protocols", Journal of Information Assurance and Security, pp. 67-79, January 2008.

[2] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, Handbook of Applied Cryptography. CRC Press, 1996.

[3] S. Rafaeli and D. Hutchison, "A Survey of Key Management for Secure Group Communication," ACM Comput. Surv., vol. 35, no. 3, pp. 309–329, 2003.

[4] H. Harney and C. Muckenhirn. "Group Key Managemant Protocol(GKMP) Architecture". July 1997, RFC 2093.

[5] H. Harney and C. Muckenhirn. "Group Key Managemant Protocol(GKMP) Specification". July 1997, RFC 2094.

[6] H.H. Chu, L. Qiao, and K. Nahrstedt. A Secure Multicast Protocol with Copyright Protection. ACM SIGCOMM Computer Communications Review, 32(2):42:60, April 2002.

[7] Y. Challal, H. Bettahar, and A. Bouabdallah. "SAKM: A Scalable and Adaptive Key Management Approach for Multicast Communications". ACM SIGCOMM Computer Communications Review, 34(2):55–70, April 2004.

[8] Bibo Jiang, Xiulin Hu, A Survey of Group Key Management, International Conference on Computer Science and Software Engineering, 2008, IEEE, DOI 10.1109/CSSE.2008.1282

[9] G. H. Chiou and W. T. Chen. Secure Broadcast using Secure Lock. IEEE Transactions on Software Engineering, 15(8):929–934, August 1989.

[10] Angelo Rossi, Samuel Pierre and Suresh Krishnan, An Efficient and Secure Self-Healing Scheme for LKH, Journal of Network and Systems Management, Vol. 18, Number 3, 327-347

[11] Debby M. Wallner, Eric J. Harder, and Ryan C. Agee. Key management for multicast: Issues and architectures. Internet draft, Network working group,september 1998, 1998.

[12] Junbeom and Hyunsoo, 2009, A decentralized multi-group key management scheme, IEICE Transactions Communications, Vol E-92-B, No.2, Feb 2009

[13] R. Canetti, T. Malkin, and K. Nissim, "Efficient Communication-Storage Tradeoffs for Multicast Encryption," in EUROCRYPT. New York, NY, USA: Springer-Verlag New York, Inc., 1999, pp. 459–474.

[14] C. K. Wong, M. Gouda, and S. S. Lam. Secure Group Communications Using Key Graphs. ACM SIGCOMM, 1998.

[15] A. Perrig, D. Song, and J.D. Tygar. ELK, A new protocol for Efficient Large-group Key distribution. IEEE Security and Priavcy Symposium, May 2001.

[16] M. Waldvogel, G. Caronni, D. Sun, N. Weiler, , and B. Plattner. The VersaKey Framework : Versatile Group Key Management, IEEE Journal on Selected Areas in Communications (Special Issues on Middleware), 17(8):1614–1631, August 1999.

[17] T.Hardjono, B.cain, and L.Monga. "Intra-domain Group key Management for Multicast Security". IETF internet Draft, September 2000

[18] A.Ballardie. Scalable "Multicast Key Distribution". May 1996, RFC1949.

[19] S. Rafaeli and D. Hutchison. Hydra: a decentralized group key management. 11th IEEE International WETICE: Enterprise Security Workshop, June 2002.

[20] Suvo Mittra, "Iolus: A Framework for Scalable Secure Multicasting", ACM SIGCOMM, 1997

[21] MOLVA, R. AND PANNETRAT, A. 1999. Scalable multicast security in dynamic groups. In Proceedings of the 6th ACMConference on Computer and Communications Security. (Singapore, Nov.). ACM, New York, 101–112.

[22] S.Setia, S.Koussih, S.Jaodia, and E.Harder. "Kronos: A scalable Group Re-Keying Approach for Secure Multicast". Proc. of IEEE Symposium on Security and Privacy,2000

[23] Y.R. Yang, X.S. Li, X.B. Zhang, and S.S. Lam. Reliable Group Rekeying: A Performance Analysis. TR-01-21, June 2001.

[24] KIM, Y., PERRIG, A., AND TSUDIK, G. 2000. Simple and fault-tolerant key agreement for dynamic collaborative groups. In Proceedings of the 7th ACM Conference in Computer and Communication Security, (Athens, Greece Nov.). (S. Jajodia and P. Samarati, Eds.), pp. 235–241.

[25] F.Staddon, S.Miner, M.Franklin, D.Balfanz, and D.Dean. "Selfhealing Key Distribution with Revocation". In Proc. Of the IEEE Symposium on Security and Privacy, Oakland, CA, May 2002.

[26] Alireza Nemaney Pour, Kazuya Kumekawa, Toshihiko Kato, Shuichi Itoh, "A Hierarchical group key management scheme for secure multicast increasing efficiency of key distribution in leave operation", Elsevier,Computer Networks, August 2007.

AUTHORS PROFILE

[1]R.Siva Ranjani is a research scholar in Andhra University under the supervision of Prof.P.S.Avadhani and Dr.D.Lalitha Bhaskari in Computer Science and Systems Engineering. She received her M.Tech (CSE) from Andhra University and presently working as Associate Professor in CSE Department of GMRIT. She is a Life Member of ISTE. Her research areas include Network Securiy, Cryptography, Group Key Managment.

[2]Mrs. Dr. D. Lalitha Bhaskari is an Associate Professor in the department of Computer Science and Engineering of Andhra University. She is guiding more than 8 Ph. D Scholars from various institutes. Her areas of interest include Theory of computation, Data Security, Image Processing, Data communications, Pattern Recognition. Apart from her regular academic activities she holds prestigious responsibilities like Associate Member in the Institute of Engineers, Member in IEEE, Associate Member in the Pentagram Research Foundation, Hyderabad, India.

[3]Dr. P. S. Avadhani is a Professor in the department of computer Science and Engineering of Andhra University. He has guided 7 Ph. D students, 3 students already submitted and right now he is guiding 12 Ph. D Scholars from various institutes. He has guided more than 100 M.Tech. Projects. He received many honors and he has been the member for many expert committees, member of Board of Studies for various universities, Resource person for various organizations. He has co-authored 4 books. He is a Life Member in CSI, AMTI, ISIAM, ISTE, YHAI and in the International Society on Education Technology. He is also a Member of IEEE, and a Member in AICTE.

CluSandra: A Framework and Algorithm for Data Stream Cluster Analysis

Jose R. Fernandez
Department of Computer Science
University of West Florida
Pensacola, FL, USA

Eman M. El-Sheikh
Department of Computer Science
University of West Florida
Pensacola, FL, USA

***Abstract*—The clustering or partitioning of a dataset's records into groups of similar records is an important aspect of knowledge discovery from datasets. A considerable amount of research has been applied to the identification of clusters in very large multi-dimensional and static datasets. However, the traditional clustering and/or pattern recognition algorithms that have resulted from this research are inefficient for clustering data streams. A data stream is a dynamic dataset that is characterized by a sequence of data records that evolves over time, has extremely fast arrival rates and is unbounded. Today, the world abounds with processes that generate high-speed evolving data streams. Examples include click streams, credit card transactions and sensor networks. The data stream's inherent characteristics present an interesting set of time and space related challenges for clustering algorithms. In particular, processing time is severely constrained and clustering algorithms must be performed in a single pass over the incoming data. This paper presents both a clustering framework and algorithm that, combined, address these challenges and allows end-users to explore and gain knowledge from evolving data streams. Our approach includes the integration of open source products that are used to control the data stream and facilitate the harnessing of knowledge from the data stream. Experimental results of testing the framework with various data streams are also discussed.**

Keywords-data stream; data mining; cluster analysis; knowledge discovery; machine learning; Cassandra database; BIRCH; CluStream; distributed systems.

I. INTRODUCTION

According to the International Data Corporation (IDC), the size of the 2006 digital universe was 0.18 zettabytes[1] and the IDC has forecasted a tenfold growth by 2011 to 1.8 zettabytes [17]. One of the main sources of this vast amount of data are streams of high speed and evolving data. Clustering analysis is a form of data mining whose application has, relatively recently, started to be applied to data streams. The unbounded and evolving nature of the data that is produced by the data stream, coupled with its varying and high-speed arrival rate, require that the data stream clustering algorithm embrace these properties: efficiency, scalability, availability, and reliability. One of the objectives of this work is to produce a distributed framework that addresses these properties and, therefore, facilitates the development of data stream clustering algorithms for this extreme environment. Another objective is to implement a clustering algorithm that is specifically designed to leverage the distributed framework. This paper describes that clustering algorithm and the distributed framework, which is entirely composed of off-the-shelf open source components. The framework is referred to simply as *CluSandra*, while the algorithm, which is deployed onto the framework, is referred to as the *CluSandra algorithm*. CluSandra's primary pillars are a database system called Cassandra [9][15] and a message queuing system (MQS). Cassandra, which is maintained by the Apache Software Foundation (ASF), is a new breed of database system that is referred to as a NoSQL database. At its core, Cassandra is a distributed hash table (DHT) designed to tackle massive datasets, perform in near-time and provide linear scalability [9]. The MQS can be any number of either open source or commercial message queuing systems that implement the Java Message Service (JMS) API. All experimentation, related to this work, was performed using the Apache ActiveMQ [16] queuing system.

The combination of the CluSandra framework and algorithm provides a distributed, scalable and highly available clustering system that operates efficiently within the severe temporal and spatial constraints associated with real-time evolving data streams. Through the use of such a system, end-users can also gain a deeper understanding of the data stream and its evolving nature in both near-time and over different time horizons.

A. Data Stream

A data stream is an ordered sequence of structured data records with these inherent characteristics: fast arrival rate, temporally ordered, evolves over time, and is unbounded [8]. The data stream's arrival rate can be in the order of thousands of data records per second, the concepts that are derived from the data stream evolve at varying rates over time and, because the data stream is unbounded, it is unfeasible to store all of its records in any form of secondary storage (e.g., DBMS). The data stream's evolutionary characteristic is referred to as *concept drift* [3]. This type of change may come as a result of the changing environment of the problem; e.g., floating probability distributions, migrating clusters of data, loss of old and appearance of new classes and/or features, class label swaps, etc. [20] Examples of data streams include IP network traffic, sensor networks, wireless networks, radio frequency identification (RFID), customer click streams, telephone records, etc. Today, there are many applications whose data is best modeled as a data stream and not as a persistent set of

[1] One zettabyte equals 10^{21} bytes or one billion terabytes.

tables. The following are some examples of applications for data stream processing [11]:

- Real-time monitoring of information systems that generate vast amounts of data. For example, computer network management, telecommunications call analysis, internet applications (e.g., Google, eBay, recommendation systems, click stream analysis) and monitoring of power plants.
- Generic software for applications based on streaming data. For example, finance (fraud detection, stock market analysis), sensor networks (e.g., environment, road traffic, weather forecasting, electric power consumption).

In this paper, a data stream S is treated as an unbounded sequence of pairs ⟨s,t⟩ , where *s* is a structured data record (set of attributes) and *t* is a system-generated timestamp attribute that specifies when the data record was created. Therefore, *t* may be viewed as the data stream's primary key and its values are monotonically increasing [7]. The timestamp values of one data stream are independent from those of any other data stream that is being processed within CluSandra. Since data streams comprise structured records, streams comprising unstructured data (e.g., audio and video streams) are not considered data streams within the context of this paper.

B. Cluster Analysis

Cluster analysis or clustering is a process by which similar objects are partitioned into groups. That is, all objects in a particular group are similar to one another, while objects in different groups are quite dissimilar. The clustering problem is formally defined as follows: *for a given set of data points, we wish to partition them into one or more groups of similar objects, where the notion of similarity is defined by a distance function [21]*. Clustering is a very broad topic that lies at the intersection of many disciplines such as statistics, machine learning, data mining, and linear algebra [12]. It is also used for many applications such as pattern recognition, fraud detection, market research, image processing, and network analysis.

The focus of this work is on *data clustering*, which is a type of data mining problem. Large multi-dimensional datasets are typically not uniformly distributed. By identifying the sparse and dense areas of the data space, data clustering uncovers the distribution patterns of the dataset [10]. In general, data clustering seeks to partition *unlabeled* data records from a large dataset into labeled clusters, which is a form of *classification*. Classification is an important problem that has been studied extensively within the context of data streams[3][4]. With respect to evolving data streams, clustering presents an attractive advantage, because it is easily adapted to changes in the data and can, therefore, be used to identify features that distinguish different clusters [12]. This is ideal for concept drifting data streams.

There are different data types (e.g., binary, numerical, discrete) that need to be taken into account by data clustering algorithms; the CluSandra algorithm only processes numerical data. Future work can deploy additional algorithms, designed to handle other data types, onto the CluSandra framework. This work also assumes that the values for all the data records' attributes are *standardized*; therefore, there is no preprocessing of the data records. Numerical data are continuous measurements of a roughly linear scale [12]. When working with data records whose attributes are of this data type, the records can be treated as n-dimensional vectors, where the similarity or dissimilarity between individual vectors is quantified by a distance measure. There are a variety of distance measures that can be applied to n-dimensional vectors; however, the most common distance measure used for continuous numerical data is the Euclidean measure:

$$d(i,j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad (1)$$

where x_{ik} and x_{jk} are the k^{th} variables for the n-dimensional data records i and j. For example, suppose you have two 2-dimenisonal data records as follows: (1,3) and (4,1). The Euclidean distance between these two records is the following:

$$\sqrt{(1-4)^2+(3-1)^2} = 3.60 \quad (2)$$

If the data stream's records are viewed as Euclidean vectors in Euclidean n-space, the distance between any two vectors (records) is the length of the line connecting the two vectors' tips or points. The lower the resulting value, the closer (similar) the two vectors. The CluSandra algorithm utilizes Euclidean distance as a measure to determine how similar or close a new data record is to a cluster's centroid (mean). It is also used to find the distance between two clusters' centroids.

The next section discusses related work in this area. Section III describes the CluSandra framework and Section IV describes the cluster query language that was developed. Section V presents the experimental results and section VI discusses the conclusions and opportunities for future work.

II. Related Work

A considerable amount of research has been applied to clustering very large multi-dimensional datasets. One of the key challenges, which has been the subject of much research, is the design of data clustering algorithms that efficiently operate within the time and space constraints presented by very large datasets. That is, the amount of available memory and required I/O time relative to the dataset's size. These constraints are greatly amplified by the data stream's extremely fast arrival time and unbounded nature. The research work done in [5], [6], and [10] introduced concepts, structures and algorithms that made great strides towards efficient clustering of data streams. The CluSandra algorithm is based on and expands on this work, as described below.

A. BIRCH

The CluSandra algorithm is based on the concepts and structures introduced by the Balanced Iterative Reducing and Clustering (BIRCH) [10] clustering algorithm. It is based on the K-means (center-based) clustering paradigm and, therefore, targets the *spherical Gaussian* cluster. K-means provides a well-defined objective function, which intuitively

coincides with the idea of clustering [19]. The simplest type of cluster is the spherical Gaussian [19]. Clusters that manifest non-spherical or arbitrary shapes, such as *correlation* and *non-linear correlation* clusters, are not addressed by the BIRCH and CluSandra algorithms. However, the CluSandra framework does not preclude the deployment of algorithms that address the non-spherical cluster types.

BIRCH mitigates the I/O costs associated with the clustering of very large multi-dimensional and persistent datasets. It is a batch algorithm that relies on multiple sequential phases of operation and is, therefore, not well-suited for data stream environments where time is severely constrained. However, BIRCH introduces concepts and a *synopsis* data structure that help address the severe time and space constraints associated with the clustering of data streams. BIRCH can typically find a good clustering with a single pass of the dataset [10], which is an absolute requirement when having to process data streams. It also introduces two structures: *cluster feature* (CF) and *cluster feature tree*. The CluSandra algorithm utilizes an extended version of the CF, which is a type of synopsis structure. The CF contains enough statistical summary information to allow for the exploration and discovery of clusters within the data stream. The information contained in the CF is used to derive these three spatial measures: *centroid*, *radius*, and *diameter*. All three are an integral part of the BIRCH and CluSandra algorithms. Given N n-dimensional data records (vectors or points) in a cluster where i = {1,2,3, …, N}, the centroid $\vec{x}_0$, radius R, and diameter D of the cluster are defined as

$$\vec{x}_0 = \frac{\sum_{i=1}^{N} \vec{x}_i}{N} \tag{3}$$

$$R = \sqrt{\frac{\sum_{i=1}^{N} (\vec{x}_i - \vec{x}_0)^2}{N}} \tag{4}$$

$$D = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} (\vec{x}_i - \vec{x}_j)^2}{N(N-1)}} \tag{5}$$

where $\vec{x}_i$ and $\vec{x}_j$ are the i[th] and j[th] data records in the cluster and N is the total number of data records in the cluster. The centroid is the cluster's mean, the radius is the average distance from the objects within the cluster to their centroid, and the diameter is the average pair-wise distance within the cluster. The radius and diameter measure the tightness or density of the objects around their cluster's centroid. The radius can also be referred to as the "root mean squared deviation" or RMSD with respect to the centroid. Typically, the first criterion for assigning a new object to a particular cluster is the new object's Euclidean distance to a cluster's centroid. That is, the new object is assigned to its closest cluster. Note, however, that equations 3, 4 and 5 require multiple passes of the data set. Like BIRCH, the CluSandra algorithm derives the radius and centroid while adhering to the single pass constraint. To accomplish this, the algorithm maintains statistical *summary* data in the CF. This data is in the form of the total number of data records (N), linear sum, and sum of the squares with respect to the elements of the n-dimensional data records. This allows the algorithm to calculate the radius, as follows:

$$\sqrt{\left(\sum x_i^2 - \left(\sum x_i\right)^2 / N\right) 1/(N-1)} \tag{6}$$

For example, given these three data records: (0,1,1), (0,5,1), and (0,9,1), the linear sum is (0,15,3), the sum of the squares is (0,107,3) and N is 3. The CF, as described in BIRCH, is a 3-tuple or triplet structure that contains the aforementioned statistical summary data. The CF represents a cluster and records summary information for that cluster. It is formally defined as

$$CF = \langle N, LS, SS \rangle \tag{7}$$

where N is the total number of objects in the cluster (i.e., the number of data records absorbed by the cluster), LS is the linear sum of the cluster and SS is the cluster's sum of the squares:

$$LS = \sum_{i=1}^{N} \vec{x}_i \tag{8}$$

$$SS = \sum_{i=1}^{N} \vec{x}_i^2 \tag{9}$$

It is interesting to note that the CF has both the *additive* and *subtractive* properties. For example, if you have two clusters with their respective CFs, CF_1 and CF_2, the CF that is formed by merging the two clusters is simply $CF_1 + CF_2$.

B. CluStream

CluStream [5] is a clustering algorithm that is specifically designed for *evolving* data streams and is based on and extends the BIRCH algorithm. This section provides a brief overview of CluStream and describes the problems and/or constraints that it addresses. This section also describes how the CluSandra algorithm builds upon and, at the same time, deviates from CluStream.

One of CluStream's goals is to address the temporal aspects of the data stream's single-pass constraint. For example, the results of applying a single-pass clustering algorithm, like BIRCH, to a data stream whose lifespan is 1 or 2 years would be dominated by outdated data. CluStream allows end-users to explore the data stream over different time horizons, which provides a better understanding of how the data stream evolves over time. CluStream divides the clustering process into two phases of operation that are meant to operate simultaneously. The first, which is referred to as the *online* phase, efficiently computes and stores data stream summary statistics in a structure called the *microcluster*. A microcluster is an extension of the BIRCH CF structure

whereby the CF is given two temporal dimensions. The second phase, which is referred to as the *offline* phase, allows end-users to perform *macroclustering* operations on a set of microclusters. Macroclustering is the process by which end-users can explore the microclusters over different time horizons. To accomplish this, CluStream uses a *tilted time frame* model for maintaining the microclusters. The tilted time frame approach stores snapshots (sets) of microclusters at different levels of granularity based on elapsed time. In other words, as time passes, the microclusters are merged into coarser snapshots. The CluSandra algorithm is based upon and extends CluStream and the design of the CluSandra framework is based on the concepts of both microclustering and macroclustering. However, the general approach taken by CluSandra for these two operational phases is quite different than that taken by CluStream.

CluStream's microcluster extends the CF structure by adding two temporal scalars or dimensions to the CF. The first scalar is the sum of the timestamps of all the data records that have been absorbed by the cluster and the second is the sum of the squares of the timestamps. Thus the CF triplet, as defined by BIRCH, is extended as follows:

$$CF = \langle N, LS, SS, ST, SST \rangle \tag{10}$$

Where ST is the sum of the timestamps and SST is the sum of the squares of the timestamps. Note that this extended CF retains its additive and subtractive properties. From henceforth, this extended version of the CF is simply referred to as a microcluster. The ST and SST can be applied to expression (6) to arrive at the temporal standard deviation of the microcluster as follows:

$$\sqrt{\left(SST - \left(ST\right)^2\right) 1/\left(N - 1\right)} \tag{11}$$

CluStream's microclustering process collects and maintains the statistical information in such a manner that the offline macroclustering phase can make effective use of the information. For example, macroclustering over different time horizons and exploring the evolution of the data stream over these horizons.

CluStream defines a fixed set of microclusters that it creates and maintains in-memory. This set, $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_q\}$ is the current online working set with q being the maximum number of microclusters in the set and each microcluster in $\mathcal{M}$ being given a unique *id*. The CluStream paper states that the algorithm is very sensitive to any further additions to the microcluster snapshot, as this negatively impacts q. This type of space constraint is one that CluSandra removes by relying on Cassandra to serve as a highly reliable and scalable real-time distributed cluster database. Like CluStream, the CluSandra algorithm also maintains a working set of microclusters in memory; however, the size of this working set is dictated by the size of a sliding time window in combination with a maximum boundary threshold (MBT).

CluStream's updating of the microclusters is similar to that of BIRCH. When a new data record is presented by the data stream, it is assigned to the closest microcluster $\mathcal{M}_i$ in $\mathcal{M}$. After finding the closest microcluster $\mathcal{M}_i$, it is then determined if $\mathcal{M}_i$ can absorb the new record without exceeding its MBT. CluStream defines the MBT as, "*a factor of t of the RMSD of the data points in* $\mathcal{M}_i$ *from the centroid*". This is another way of referring to the standard deviation of the cluster times a factor t to arrive at the final maximum radius. The microcluster $\mathcal{M}_i$ is updated whenever it absorbs a new data record. If $\mathcal{M}_i$ has absorbed only one data record, the RMSD cannot be calculated; therefore, for those clusters having absorbed only one data record, the MBT is derived as the distance from $\mathcal{M}_i$ to its closest neighbor times a factor r. The CluSandra algorithm uses a similar approach in locating the closest microcluster. However, it does not use the nearest neighbor approach for those instances where the closest microcluster has absorbed only one entry. It instead determines if the closest microcluster can absorb the data record without exceeding a configurable and fixed maximum radius, which is CluSandra's MBT.

If $\mathcal{M}_i$ cannot absorb the new data record, CluStream creates a new microcluster to host the data record. To conserve memory, CluStream requires that either one of the existing microclusters in $\mathcal{M}$ be deleted or two microclusters be merged. Only those microclusters that are determined to be *outliers* are removed from $\mathcal{M}$. If an outlier cannot be found in $\mathcal{M}$, two microclusters are merged. The CluSandra algorithm deviates from this approach since it simply adds a new microcluster to its current working set. Again, the size of CluSandra's in-memory working set is managed according to a temporal sliding window and the specified MBT. Any microclusters that are no longer active within the current sliding window are removed from the working set, but not before being persisted to the Cassandra cluster database.

The temporal scalars (ST, SST) of the microcluster, in combination with a user-specified threshold δ, are used to look for an outlier microcluster in $\mathcal{M}$. The ST and SST scalars allows CluStream to calculate the mean and standard deviation of the arrival times of the data records in $\mathcal{M}$'s microclusters. CluStream assumes that the arrival times adhere to a normal distribution. With the mean and standard deviation of the arrival times calculated, CluStream calculates a "*relevance stamp*" for each of the microclusters. A microcluster whose relevance stamp is less than the threshold δ is considered an outlier and subject to removal from $\mathcal{M}$. If all the relevance stamps are recent enough, then it is most likely that there will be no microclusters in $\mathcal{M}$ whose relevance stamp is less than δ. If and when this occurs, CluStream merges the two closest microclusters in $\mathcal{M}$ and assigns the resulting merged microcluster a *listid* that is used to identify the clusters that were merged to create this new merged microcluster. So as time progresses, one microcluster may end up comprising many individual microclusters.

Unlike BIRCH, CluStream does not utilize a tree structure to maintain its microclusters. At certain time intervals, and while $\mathcal{M}$ is being maintained as described above, $\mathcal{M}$ is persisted to secondary storage. Each instance of $\mathcal{M}$ that is persisted is referred to as a *snapshot*. CluStream employs a logarithmic based time interval scheme, which is referred to as

a *pyramidal time frame*, to store the snapshots. This technique guarantees that all individual microclusters in $\mathcal{M}$ are persisted prior to removal from $\mathcal{M}$ or being merged with another microcluster in $\mathcal{M}$. This allows a persisted and merged microcluster (i.e., those having a listid) in a snapshot to be broken down (via the microclusters subtractive property) into its constituent (individual), finer-grained microclusters during the macroclustering portion of the process. The opposite is also available, whereby the additive property allows finer-grained/individual and merged microclusters to be merged into more course grained microclusters that cover specified time horizons. Snapshots are classified into orders, which can vary from 1 to $\log_2(T)$, where T is the amount of clock time elapsed since the beginning of the stream[5]. The number of snapshots stored over a period of T time units is

$$(a+1)*\log_2(T) \tag{12}$$

For example, if $\alpha = 2$ and the time unit or granularity is 1 second, then the number of snapshots maintained over 100 years is as follows:

$$(2+1)*\log_2(100*365*24*60*60) \gg 95 \tag{13}$$

The CluSandra algorithm does not operate within the same memory constraints as CluStream. During the CluSandra algorithm's microclustering process, microclusters are not merged to accommodate a new microcluster and there is no need to search for possible outliers that can be targeted for removal from $\mathcal{M}$. When a new microcluster is created, it is simply added to the current in-memory working set and persisted to the Cassandra cluster database. Also, when a microcluster absorbs a data record, the microcluster is immediately persisted to the cluster database; there is no dependence on a periodic time interval scheme for persisting $\mathcal{M}$ to secondary storage.

CluStream does not include a database system of any kind for persisting its snapshots and does not fully address the transactional integrity associated with snapshot persistence. After the in-memory snapshot $\mathcal{M}$ has been updated, there may exist a relatively lengthy time period before the snapshot is persisted to the local file system. This raises the risk that the snapshot's state may be lost due to process, system or hardware failure. Also, CluStream does not address $\mathcal{M}$'s recoverability. If the machine fails and is restarted, CluStream cannot reliably recover $\mathcal{M}$'s state prior to the failure. In other words, there is no transactional integrity associated with $\mathcal{M}$'s persistence. One of the goals of the CluSandra framework is to introduce the necessary components that guarantee the transactional integrity of microcluster persistence.

III. CluSandra Framework

Figure 1 is a level 1 context-level data flow diagram[2] (DFD) that represents, at a high-level, the CluSandra framework. The framework's core components are written entirely in version 1.5 of the Java programming language; however, the framework supports client components that are written in a variety of programming languages. The CluSandra framework and algorithm are based on temporal and spatial aspects of clustering. The algorithm, which is implemented in a MicroClustering Agent (MCA) framework component and described in more detail in section D, uses temporal and spatial measures (radius, distance) to group the data stream's records into microclusters. The microclusters are stored in the Cassandra database and later accessed by offline processes (e.g., aggregator) and/or end-users. A Cluster Query Language (CQL) is provided by the framework to facilitate the querying and analysis of the microclusters in the database. As previously noted, Cassandra is an implementation of a DHT that comprises two or more distributed machines configured in a peer-to-peer ring network topology. The ring of machines or nodes is called a Cassandra cluster. To meet the most demanding environments, Cassandra can *elastically* scale up from a one or two node cluster to a cluster comprising 10s, if not 100s, of nodes and then back down to one or two nodes. Each node in the cluster may also optionally host the CluSandra framework's other executable components.

A. StreamReader

The StreamReader component is responsible for reading the data stream's structured data records, wrapping those data records in a CluSandra-specific DataRecord object, and then sending the DataRecords to the CluSandra message queuing system (MQS), where they are temporarily stored or buffered for subsequent processing. The CluSandra framework automatically time stamps the DataRecords when they are created, but the StreamReader can also override the framework's timestamp. This may be required if, for example, the raw data stream records are already time-stamped. The time stamps are critically important, because they are used to record the data stream's timeline. The raw data stream record that is read by the StreamReader is treated as a multidimensional vector containing one or more continuous numerical values. This vector is encapsulated by the DataRecord object. It is critical for the StreamReader to keep up with the data stream's arrival rate, which is assumed to be in the 1000s of records per second[2]. However, this should not be an issue given the simple and straightforward nature of the StreamReader's main purpose.

[2] It is also quite possible that the StreamReader is the stream generator.

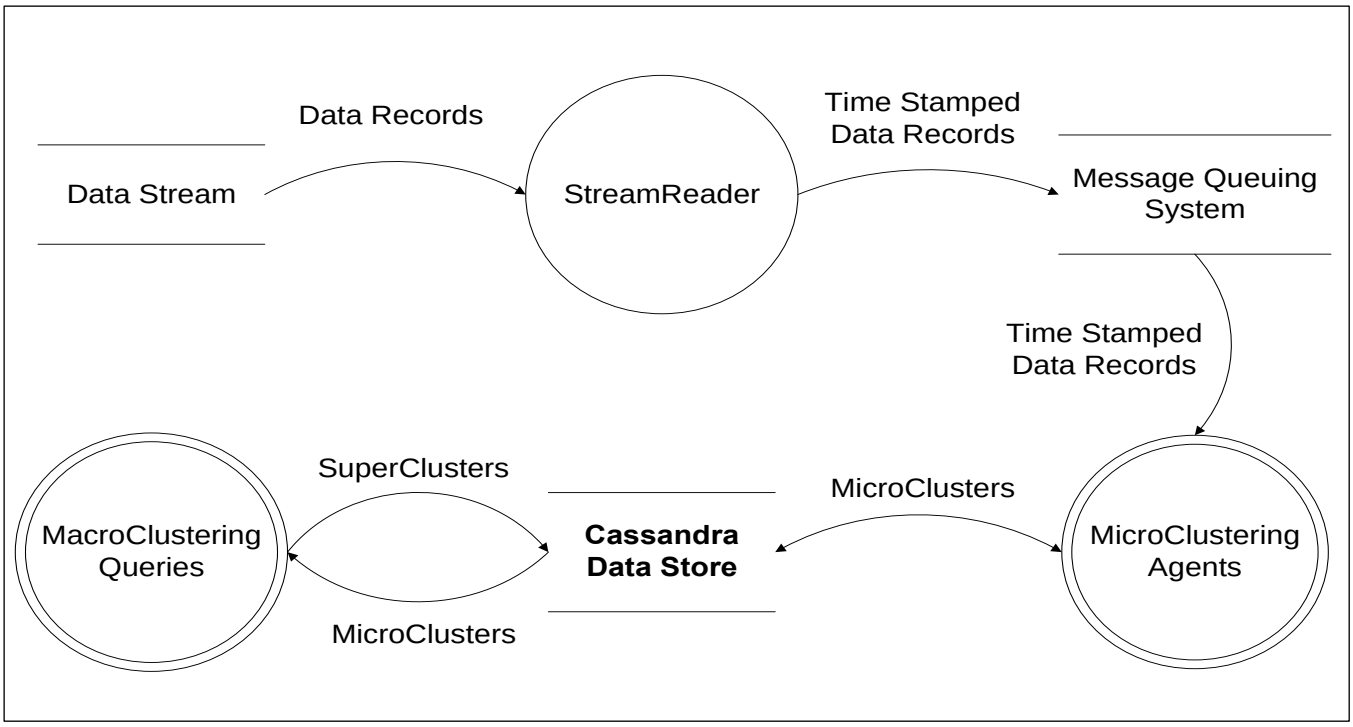


Figure 1. CluSandra data flow diagram

CluSandra's MQS provider is remotely accessed via a TCP/IP network; therefore, the StreamReader component does not have to reside on the same node as the MQS provider nor does it have to reside in a Cassandra cluster node. The StreamReader can even be embedded within the component or device (e.g., sensor or router) that produces the data stream. StreamReaders can be distributed across a population of such devices, all writing to the same or different MQS queue, with each queue dedicated to a particular data stream. Most MQS providers, like ActiveMQ, support clients written in a variety of languages. If a StreamReader is not written in the Java programming language and would like to take advantage of the component that implements the CluSandra algorithm, then it requires a transformational component; in other words, a Java proxy that can create a Java DataRecord object and place it in the MQS for the non-Java StreamReader.

B. Timeline Index

The timeline index (TI), which is not depicted in the DFD, represents the data stream's timeline. It is an important temporal-based index that is maintained in the Cassandra database and used for maintaining the clusters in the database. In the Cassandra vernacular, the TI is a type of *secondary index*. There is one TI defined for each data stream that is being processed within the CluSandra framework. The TI is implemented as a Cassandra *SuperColumnFamily*. At an abstract level, each entry in the TI represents a second in the data stream's lifespan. Each entry's contents or value is a Cassandra *SuperColumn* whose entries (Columns) form a collection or set of row keys to clusters that were or are still active during that second in time. A data stream's timeline can run for any amount of time. During development and testing, the stream's timeline may extend over a handful of seconds, while in production environments, the timeline may extend over months, if not years.

The TI's implementation comprises a Cluster Index Table (CIT). The CIT includes a row for each day of the year and each row comprises 86400 SuperColumns; one SuperColumn for each second of that particular day. Each SuperColumn contains one or more Columns whose values are keys to clusters in the Cluster Table (CT). Also, a SuperColumn can be assigned different names from row-to-row, and in this case the name of a SuperColumn is a timestamp that corresponds to a second for that day. Each row key of the CIT is given a value that corresponds to the zeroth second of a particular day. The motivation behind the TI's implementation is that Cassandra is not currently set up to perform well with sorted rows and returning ranges within those sorted rows. However, it is set up to sort columns and return ranges or slices of columns, based on their names.

C. Message Queuing System

The message queuing system (MQS) is a critical piece of the CluSandra framework. It serves as a reliable asynchronous message store (buffer) that guarantees delivery of its queued messages (DataRecords). So as a dam is used to control a wild raging river and harness it to produce electricity, so is the MQS used to control the evolving high-speed data stream so that it can be harnessed to produce knowledge. In CluSandra's case, a MQS queue serves as the stream's dam and its contents of time-stamped DataRecords form the reservoir. The CluSandra framework supports the simultaneous processing of many data streams; therefore, there may very well be many queues defined within the MQS; one queue for each data stream.

The primary motivation for having the CluSandra framework incorporate a MQS is to *control the data stream*. More precisely, the MQS provides a reliable DataRecord store that temporarily buffers and automatically distributes DataRecords across one or more instances of the microclustering agent (MCA) component. A group of identical MCAs that consume DataRecords from the same queue is called a *swarm*. Many swarms can be defined and distributed across the framework, with each swarm reading from its unique queue. The queue is capable of retaining thousands of DataRecords and guarantees their delivery to the swarm, which is responsible for ensuring that the number of DataRecords in the MQS queue is maintained at an acceptable level. The swarm size is, therefore, a function of the data stream rate. Data streams with extremely fast arrival rates produce very large volumes of DataRecords and require a correspondingly large swarm.

The CluSandra framework is designed to utilize any MQS that implements the Java Message Service (JMS) API. The vast majority of MQS providers, both open source and commercial, implement the JMS. The JMS is an industry standard Java messaging interface that decouples applications, like the StreamReader, from the different JMS implementations. JMS thus facilitates the seamless porting of JMS-based applications from one JMS-based queuing system to another. Within the context of the JMS, the process that places messages in the MQS's queues is called the *producer*, the process that reads messages from the queue is called the *consumer*, and both are generically referred to as *clients*. So the StreamReader is a producer, the MCA is a consumer and they are both clients.

The MQS *guarantees* delivery of DataRecords to the swarm. This guarantee means that DataRecords are delivered even if the MQS or machine hosting the MQS were to fail and be restarted. The MQS achieves this guarantee via a combination of message persistence and a broker-to-client acknowledgment protocol. These MQSs can be configured to persist their messages to either file systems (distributed or local) or database management systems (DBMS). For the CluSandra framework, the MQS is configured to use a distributed or shared file system and not a DBMS. The file system provides much better throughput performance than does a DBMS.

These MQS systems are also architected to provide fault-tolerance via redundancy. For example, you can run multiple "*message broker*" processes across multiple machines, where certain message brokers can act as hot or passive standbys for failover purposes. The message broker is the core component of the MQS and is the component responsible for message delivery. One overarching requirement is to ensure that this level of reliability and/or fault tolerance be an inherent quality of the CluSandra framework. These MQS systems include many features, but it is beyond the scope of this paper to list all the features.

D. Microclustering Agent

The microclustering agent (MCA) is the framework component that consumes DataRecords from a particular MQS queue and implements a clustering algorithm that produces microclusters. This section describes the MCA that implements the CluSandra algorithm and is delivered with the CluSandra framework.

Like CluStream, the CluSandra algorithm tackles the one-pass data stream constraint by dividing the data stream clustering process into two operating phases: online and offline. Microclustering takes place in real-time, computing and storing summary statistics about the data stream in microclusters. There is also an optional offline aggregation phase of microclustering that merges temporally and spatially similar microclusters. Macroclustering is another offline process by which end-users create and submit queries against the stored microclusters to discover, explore and learn from the evolving data stream. As CluStream extended the BIRCH CF data structure, so does the CluSandra algorithm extend the CluStream's CF structure as follows:

$$CFT = \langle N, LS, SS, ST, SST, CT, LAT, IDLIST \rangle \qquad (14)$$

The CFT is used to represent either a microcluster or supercluster in the CluSandra data store. The term *cluster* applies to both micro and superclusters. The CT and LAT parameters are two timestamp scalars that specify the creation time and last absorption time of the cluster, respectively. More precisely, CT records the time the cluster was created and the LAT records the timestamp associated with the last DataRecord that the cluster absorbed. When a microcluster is first created, to absorb a DataRecord that no other existing microcluster can absorb, both the CT and LAT parameters are assigned the value of the DataRecord's timestamp. All timestamps in CluSandra are measured in the number of milliseconds that have elapsed since Unix or Posix epoch time, which is January 1,1970 00:00:00 GMT. The IDLIST is also new and is used by superclusters.

Over time, a particular pattern in the data stream may appear, disappear and then reappear. This is reflected or captured by two microclusters with identical or very similar spatial values, but different temporal values. The time horizon over which a cluster was active can be calculated by the CT and LAT parameters and more detailed statistical analysis can be performed based on the other temporal, as well as spatial parameters. For example, the temporal density of the DataRecords and their spatial density with respect to one another and/or their centroid. An *inactive* microcluster is no longer capable of absorbing data records; however, during macroclustering, it can be merged with other inactive and active microclusters to form a supercluster. The algorithm's microclustering phase, therefore, works within a specified temporal sliding window and updates only those microclusters that are within that time window. Such a temporal sliding window is required, because of the unbounded nature of the data stream.

As previously mentioned, the MCA consumes sets of DataRecords from its assigned MQS queue. The framework provides a *read template* for the MCA that includes this functionality; therefore, the one implementing the MCA's clustering algorithm does not need to concern herself with this functionality. The members of a set $\mathcal{D}$ of DataRecords are consumed, by the read template, in the same order that they

were produced by the StreamReader and the temporal order of the DataRecords is maintained by the MQS. The set $\mathcal{D}$ can, therefore, be viewed as a temporal window of the data stream.

$$\mathcal{D} = \{d_1, d_2, \ldots, d_k\} \quad (15)$$

The maximum number of DataRecords in $\mathcal{D}$ is configurable. Also, the amount of time the read template blocks on a queue, waiting to read the maximum number of DataRecords in $\mathcal{D}$, is configurable. The read template processes the set $\mathcal{D}$ whenever the maximum number has been reached or the read time expires and $\mathcal{D} \neq \emptyset$. The read time should be kept at a relatively high value. So, $0 < k <= m$, where k is the total number of DataRecords in $\mathcal{D}$ and m is the maximum. If the read time expires and $\mathcal{D} = \emptyset$, the MCA simply goes back to blocking on the queue for the specified read time.

After the read template has read a set $\mathcal{D}$ from the queue, it gives it to the clustering algorithm via a specified interface. The following describes the CluSandra clustering algorithm and from henceforth it is simply referred to as the CA. When the CA receives a set $\mathcal{D}$, it begins the process of partitioning the DataRecords in $\mathcal{D}$ into a set $\mathcal{M}$ of currently active microclusters. On startup, $\mathcal{M}$ is an empty set, but as time progresses, $\mathcal{M}$ is populated with microclusters. The CA then determines the time horizon h associated with $\mathcal{D}$. The range of h is R_h and it is defined by the newest and oldest DataRecords in $\mathcal{D}$. Therefore, $R_h = \{t_o\text{-}t_e, t_y\}$, where t_e is the configurable microcluster expiry time and t_o and t_y represent the oldest and newest DataRecords in $\mathcal{D}$, respectively. All microclusters in $\mathcal{M}$, whose LAT does not fall within R_h, are considered inactive microclusters and removed from $\mathcal{M}$. Thus the CA always works within a sliding temporal window that is defined by R_h. When the CA completes the processing of $\mathcal{D}$, it writes all new and updated microclusters in $\mathcal{M}$ to the Cassandra data store, gives control back to read template, and starts the partitioning process over again when it is given a new set $\mathcal{D}$ of DataRecords.

To partition the DataRecords in $\mathcal{D}$, the CA iterates through each DataRecord in $\mathcal{D}$ and selects a subset $\mathcal{S}$ of microclusters from $\mathcal{M}$, where all microclusters in $\mathcal{S}$ are active based on the current DataRecord's (d_c) timestamp. For example, if the timestamp for d_c is t_d, then only those microclusters in $\mathcal{M}$ whose LAT value falls within the range, $\{t_d\text{-}t_e, t_d\}$ are added to $\mathcal{S}$. The value t_e is the configurable microcluster expiry time. Depending on the rate of the data stream and the microcluster expiry time, it is possible that $\mathcal{S} = \mathcal{M}$ and so, $\mathcal{S} \subseteq \mathcal{M}$. The CA uses the Euclidean distance measure to find the microcluster in $\mathcal{S}$ that is closest to d_c. The MBT is then used to determine if the closest microcluster $\mathcal{M}_i$ in $\mathcal{S}$ can absorb d_c. The MBT is a configurable numeric value that specifies the maximum radius of a microcluster. Again, the radius or RMSD is the root mean squared deviation of the cluster and is derived according to (6). If $\mathcal{M}_i$ can absorb d_c without breaching the MBT, $\mathcal{M}_i$ is allowed to absorb d_c and is placed back in $\mathcal{M}$, else a new microcluster is created to absorb d_c and that new microcluster is then added to $\mathcal{M}$. If $\mathcal{S} = \emptyset$, the CA simply creates a new microcluster to absorb d_c and adds the new microcluster to $\mathcal{M}$.

One method for deriving a MBT value for the target data stream is by sampling the data stream to derive an average distance (measure of density) between DataRecords and then using some fraction of that average density. If the CA has not been assigned a MBT, it will derive the MBT based on the first set of data records that it receives from the read template. If $\mathcal{M}_i$ has previously absorbed only one DataRecord (i.e., N=1), the RMSD cannot be calculated. In such a case, the MBT is used to determine if the microcluster can absorb the DataRecord.

Depending on the distribution of the data stream, as it evolves over time, microclusters of all sizes appear, disappear, and may reappear. The number and sizes of the microclusters are a factor of not only the data stream's evolutionary pattern, but also of the MBT and microcluster expiry-time. The smaller the MBT, the more microclusters will be produced and vice versa. However, the optional offline aggregation and/or macroclustering phases can be used to merge those microclusters that are deemed similar. The next section describes the CluSandra Aggregator component, which is responsible for the aggregation and merges those microclusters whose radii overlap. Please note that the Aggregator does not produce superclusters; it simply *merges* overlapping microclusters.

E. Aggregating Microclusters

If there is an MCA swarm distributed across the CluSandra framework's nodes, then it is very likely for the swarm to create microclusters that are very similar, if not equal, both temporally and spatially (see figure 2). This occurs if two or more MCAs in the swarm process a set of DataRecords with equal or overlapping time horizons (h_e). This may also occur as a natural side effect of the clustering algorithm.

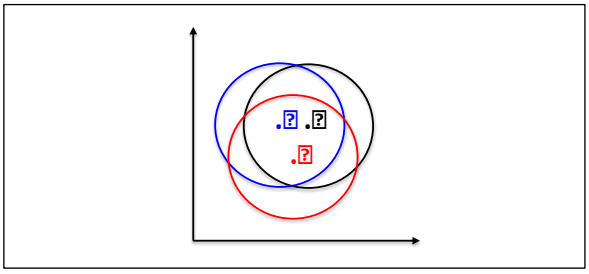


Figure 2. Overlapping clusters

If any microclusters temporally and spatially overlap, then those microclusters may be viewed as one microcluster. Given a two-dimensional vector-space, the figure above illustrates an example where three MCAs have created three microclusters (dots represent the microclusters' centroids) that are so close to one another, both spatially and temporally, that they should be merged into one microcluster. The merging of these microclusters is performed by an offline aggregator component that sweeps through the data stream timeline (CIT) performing such merges. Microcluster aggregation is a type of *agglomerative* clustering procedure whereby individuals or

groups of individuals are merged based upon their temporal and spatial proximity to one another [13]. Agglomerative procedures are probably the most widely used of the hierarchical clustering methods [13]. The result of a merger, performed by the aggregator, is immutable (see figure 3). This aggregator, which is provided as part of the CluSandra algorithm package, should not be confused with superclustering and macroclustering, which is discussed in the following section.

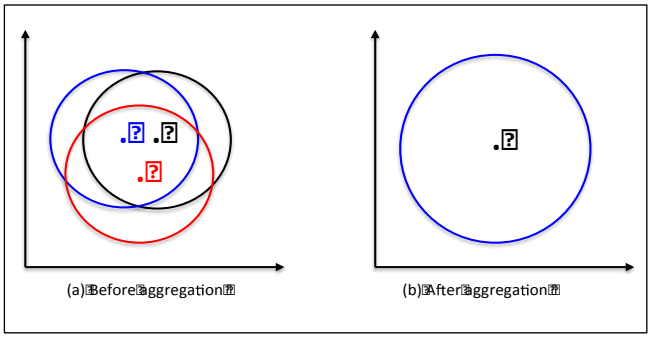


Figure 3. Before and after aggregation

The microcluster expiry time is used to determine if two microclusters temporally overlap. To determine if two temporally overlapping microclusters also spatially overlap, the aggregator compares the distance between their centroids with the sum of their radii. If the distance is less than the sum of the radii, then the two microclusters spatially overlap. There may even be instances where one microcluster is entirely within the other. The aggregator is given a configurable property called the *overlapFactor*. This property is used to specify the amount of overlap that is required to deem two microclusters similar enough to merge. The radii of the microcluster that results from the merge may be greater than the radii of the two merged microclusters; therefore, it will occupy more space, as well as time and be capable of absorbing more surrounding microclusters that temporally overlap. So the less the overlapFactor, the greater the probability of creating very large clusters that may mask out interesting patterns in the data stream.

To minimize the occurrence of overlapping microclusters, all members of a MCA swarm can work from the same or shared set of microclusters in the data store. However, this would have required coordination between the distributed members of the swarm. Unfortunately, this level of coordination requires a distributed locking mechanism that introduces severe contention between the members of the swarm. It is for this reason, that this approach of sharing microclusters was not followed. Also, overlapping microclusters is a natural side effect of the CluSandra clustering algorithm.

F. Superclusters and Macroclustering

The CluSandra framework introduces the supercluster, which is created when two or more clusters, either micro or super, are merged via their additive properties. A supercluster can also be reduced or even eliminated via its subtractive property. The IDLIST in the CluSandra CFT is a collection or vector of microcluster ids (Cassandra row keys) that identify a supercluster's constituent parts (i.e., microclusters). If the IDLIST is empty, then it identifies the cluster as a microcluster, else a supercluster.

Superclusters are created by the end-user during the macroclustering process. Superclusters are created based on a specified distance measure (similarity) and time horizon. Like aggregation, superclustering is another type of *agglomerative* clustering procedure; however, unlike traditional agglomertive procedures where the result of a merger is immutable, CluSandra's superclusters can be undone. What makes this possible is the subtractive property of the CFT. When a supercluster is created, the earliest CT and LAT of its constituent parts are used as the supercluster's CT and LAT. The CT and LAT thus identify the supercluster's lifespan.

G. The Spring Framework

The open source Spring Framework [1] is relied upon by the CluSandra framework for configuration and to reuse Spring's components. Except for Cassandra, all CluSandra framework components are comprised of one or more Spring POJOs (Plain Old Java Objects) that are configured via Spring XML configuration files. The Spring Framework, which is henceforth referred to simply as *Spring*, was created for the specific purpose of minimizing the development costs associated with Java application development. Spring provides a number of Java packages whose components are meant to be reused and that, in general, hide the complexities of Java application development. However, at its core, Spring is a modular dependency injection and aspect-oriented container and framework. By using Spring, Java developers benefit in terms of simplicity, testability and loose coupling.

The StreamReader and MCA components make heavy use of Spring's support for the JMS to send and receive DataRecords to and from the MQS, respectively. Apache ActiveMQ [16] has been selected as the CluSandra framework's default JMS provider (i.e., MQS). There are many open source JMS providers or implementations. ActiveMQ was chosen, because of its rich functionality, beyond that specified by the JMS specification, and its robust support for and integration with Spring.

As a JMS producer, the StreamReader indirectly uses Spring's JmsTemplate class to produce DataRecords destined for the ActiveMQ message broker. The JmsTemplate is based on the template design pattern and it is a convenience class that hides much of the complexity of sending messages to a JMS message broker. The JmsTemplate is integrated into a *send template* (convenience class) provided by CluSandra; therefore, the person implementing the StreamReader does not have to concern herself with the implementation of this functionality. This send template is similar to the previously described *read template* for the MCA. In general, the use of the JmsTemplate by both the send and read templates, in combination with the corresponding Spring XML file, helps ensure a decoupling between the CluSandra JMS clients and whatever JMS provider is being used as the CluSandra MQS.

IV. Cluster Query Language

The CluSandra framework includes a query language that is used for querying the CluSandra algorithm's cluster

database. The query language, which is referred to as the *cluster query language* or *CQL* for short, includes the following statements:

- *Connect*: This statement is used to connect to a particular node in the Cassandra cluster.
- *Use*: This statement, which must be invoked immediately after the Connect, is used to specify the keyspace to use within the Cassandra cluster. A Cassandra keyspace is analogous to a schema that is found in a relational database.
- *Select*: This statement, which is the one most often invoked, is used for projecting clusters from the cluster store.
- *Aggregate*: This statement is used for invoking the aggregator on all or a portion of the cluster database.
- *Merge*: This statement is used, as part of the offline macroclustering process, to form superclusters.
- *Sum*: This is a relatively simple statement that is used to return the total number of DataRecords that have been absorbed by all the microclusters in the cluster database.
- *Distance*: This statement is used for acquiring the distance between pairs of clusters.
- *Overlap*: This statement is used for acquiring the amount of overlap between pairs of clusters or in other words, the overlap percentage of the clusters' radii.

CluSandra's CQL is not to be confused with the Cassandra Query Language, which is also referred to as CQL. CluSandra's CQL operates in either batch or interactive mode. When invoked in batch mode, the user specifies a file that contains CQL statements.

V. Empirical Results

Several experiments were conducted to evaluate the accuracy of the CluSandra algorithm, as well as the scalability and reliability provided by the CluSandra framework. The experiments are considered small-scale where the framework comprised a cluster of 2-3 compute nodes. More large-scale testing that comprises larger clusters of compute nodes is planned for future work.

A. Test Environment and Datasets

Experiments designed to test the accuracy of the CluSandra algorithm were conducted on an Intel Core i5 with 8 GB of memory running OS/X version 10.6.8. The experiments included a real and synthetic dataset.

The real dataset was acquired from the 1998 DARPA Intrusion Detection Evaluation Program, which was prepared and managed by MIT Lincoln Labs. This is the same dataset used for The Third International Knowledge Discovery and Data Mining Tools Competition (KDD-99 Cup). According to [5], this dataset was also used to run experiments on the CluStream algorithm. The dataset is contained in a comma-separated values (CSV) file where each line comprises a TCP connection record; there are 4,898,431 records in the file. The dataset records two week's worth of normal network traffic, along with bursts of different types of intrusion attacks, that was simulated for a fictitious US military base. The attacks fall into four main categories: DOS (denial-of-service), R2L (unauthorized access from a remote machine), U2R (unauthorized access to local super user privileges), and PROBING (surveillance and other probing). Each connection record contains 42 attributes, which provide information regarding the individual TCP connection between hosts inside and outside the fictitious military base. Some example attributes are protocol (e.g., Telnet, Finger, HTTP, FTP, SMTP, etc), duration of the connection, the number of root accesses, number of bytes transmitted to and from source and destination. The connection records are not time stamped. Of the 43 attributes, 34 are of type continuous numerical. Every record in the dataset is labeled as either a *normal* connection or a connection associated with a particular attack. The following is a list of all possible labels, with the number in parenthesis being the total number of records in the dataset having that particular label: *back*(2203), *buffer_overflow*(30), *ftp_write*(8), *guess_passwd*(53), *imap*(1069), *ipsweep*(12481), *land*(21), *loadmodule*(9), *multihop*(7), *neptune*(1072017), *nmap*(2316), *normal*(972781), *perl*(3), *phf*(4), *pod*(264), *portsweep*(10413), *rootkit*(10), *satan*(15892), *smurf*(2807886), *spy*(2), *teardrop*(979), *warezclient*(1020), *warezmaster*(20). Together, the normal, smurf and neptune records comprise 99% of all records. The StreamReader (KddStreamReader) for this experiment reads each connection record and creates a DataRecord that encapsulates the 34 numerical attributes for that record. The KddStreamReader then sends the DataRecord the framework's MQS for processing by the MCA, which in this case is an implementation of the CluSandra algorithm. The KddStreamReader includes a filter that is configured to read all or any combination of records based on the label type. For example, the end-user can configure the KddStreamReader to process only the neptune records, only the smurf and neptune records, or all the records.

The first series of experiments focused on processing the more ubiquitous record types in the dataset. The first experiment in the series had the KddStreamReader process only the smurf records and assigned the MCA a time window that captured the entire stream and an MBT (i.e., maximum radius) of 1000; this same time window and MBT were maintained throughout this series of experiments. The result was one microcluster that had absorbed all 2,807,886 smurf records and had a relatively dense radius of 242.34. This experiment was executed five times with identical results. The second experiment was identical to the first, except that the KddStreamReader processed only the neptune records, which is the second most ubiquitous record type. The result was, once again, one microcluster that had absorbed all 1,072,017 neptune records, but with an even smaller radius of 103.15. The next experiment targeted the normal record, which is the third most ubiquitous record. One might expect that the result would, once again, be only one microcluster. However, the result was a set of 1,048 microclusters with very little to no overlap between the microclusters and a high degree of variance with respect to their radii and number of absorbed DataRecords. This relatively large set of microclusters is to be expected, because there exists a high degree of variance within a set of normal connections. In other words, in a TCP network, the usage across a set of normal connections is typically not the same; the connections are being used for a variety of different reasons. For example, some connections are being used for email (SMTP), file transfer (FTP), and

terminal interfaces (Telnet, HTTP). It was also noted that it took appreciably longer for this experiment to complete. This is due to the specified time window encompassing the data stream's entire lifespan, which results in a much larger number of microclusters in the in-memory working set. The StreamReader's next target was the ipsweep records. The result was one microcluster absorbing all but one ipsweep record; the radius of this one microcluster was 145.66. The distance between this microcluster and the one that had absorbed the remaining ipsweep record was 53,340. It was, therefore, assumed that this one neptune record was an outlier. Finally, the KddStreamReader was focused on the portsweep records. Unlike the previous attack-related experiments, where only one or two microclusters were produced, this one resulted in 24 microclusters; however, one microcluster absorbed 90% of all the portsweep records and it had a radius of 569. The other 23 microclusters had radii ranging in the 900-1000 range with very little overlap. Overall, the results of this first series of experiments indicated a high level of accuracy for the CluSandra algorithm. Also, in this series of experiments, the average throughput rate for the KddStreamReader was approximately 28,000 records per second.

In the next series of experiments, the KddStreamReader processed combinations of two or more connection record types. To start, the KddStreamReader processed only the neptune and smurf records, along with the same time window and MBT as the previous series of experiments. The result was two microclusters; the first having absorbed 2,807,639 records with a radius of 242.17, while the second absorbed 1,072,264 records with a radius of 101.52. It was clear that the first and second microcluster accurately grouped the smurf and neptune records, respectively. There was also no overlap between the two microclusters, but it was interesting to see that a relatively small number of smurf records were grouped into the neptune microcluster. The KddStreamReader was then configured to process all smurf, neptune and normal records. The result was a set of 1,053 microclusters. The most populated microcluster had absorbed 2,804,465 DataRecords and had a radius of 232.62. This is clearly the smurf microcluster, but note how it had lost approximately 3000 DataRecords. The second most populated microcluster had absorbed 1,438,988 DataRecords and had a radius of 497.34. This is clearly the neptune microcluster, but it had gained over 300,000 DataRecords and its radius had, as would be expected, appreciably increased from 101.52. It was concluded that a considerable number of normal DataRecords had been absorbed by the neptune microcluster. This can be addressed by reducing the MBT; however, this will also result in an appreciably larger number of denser normal microclusters. The third most populated microcluster had absorbed 140,885 DataRecords with a radius of 544.89. Population-wise, this is approximately one-tenth the size of the neptune microcluster.

To test the accuracy of the sliding time window, the window's size was reduced to capture the different bursts of attacks. It was noted from the analysis of the dataset file that the neptune and smurf attacks occur across a handful of different bursts. So the size of the sliding time window was reduced to 3 seconds, the MBT was kept at 1000, and the StreamReader processed only the neptune records. The result was 4 very dense microclusters (attacks) that had a radius of 100 or less, with 18 being the smallest radius. The first microcluster absorbed 15 DataRecords and its lifespan (duration) was only one second. The second microcluster was created 4 seconds later, absorbed approximately 411,000 DataRecords and its lifespan was 12 seconds. Thus there was a gap of 4 seconds, between the first and second microcluster, where there were no neptune attacks. The third microcluster was created 15 seconds after the second expired, absorbed approximately 450,000 DataRecords, and its lifespan was 9 seconds. Finally, the fourth microcluster was created 4 seconds after the third expired, absorbed approximately 211,000 DataRecords and its lifespan was 6 seconds. The test was repeated for the smurf connection records and the result was one microcluster that had absorbed all the DataRecords and had a lifespan of over one minute. The experiment was run again, but with a window of 2 seconds. This time, the result was 5 very dense microclusters; three had a lifespan of 1 or 2 seconds, one a lifespan of 12 seconds and the last a lifespan of 42 seconds. These experiments proved the accuracy of the sliding window.

The synthetic dataset was generated by a stream generator that is loosely based on the Radial Basis Function (RBF) stream generator that is found in the University of Waikato's Massive Online Analysis (MOA) open source Java package [8]. This RBF type of generator was used, because it produces data streams whose data distribution adhere to a spherical Gaussian distribution, which is the distribution that the CluSandra algorithm is designed to process. During its initialization, the RBF generator creates a set of randomly generated centroids. The number of centroids in the set is specified by one of the generator's configurable parameters. Each centroid, which represents a distinct class, is given a random standard deviation and a multivariate center that is a proper distance from all the other centroids' centers. Not ensuring a proper distance between centroids leads to ambiguous results, because the resulting radial fields associated with two or more centroids may overlap. In some cases, the amount of overlap is considerable. The number of variables or attributes assigned to the centroids' centers is also specified via a configurable parameter. A new data record is generated by first randomly selecting one of the centroids. Then a random offset with direction is created from the chosen centroid's center. The magnitude of the offset is randomly drawn from a Gaussian distribution in combination with the centroid's standard deviation. This effectively creates a normally distributed hypersphere of data records, with distinct density, around the corresponding centroid [14].

The first series of experiments, with the synthetic RBF generator, configured the generator to produce a data stream comprising 200,000 data records with 5 classes and whose records had five attributes. The MBT was set to 3.0 and the sliding time window captured the entire data stream. On some occasions, the result was as expected; i.e., five very dense microclusters with radii ranging from 0.18 to 1.4 and whose population of absorbed data records was rather evenly distributed. On other occasions, the result was more than five

microclusters. However, on these occasions, there were always five dense microclusters that had no overlap and absorbed the vast majority of the data stream's records. Of those five, there was always one that had absorbed substantially less data records than the other four. Using the CQL, it was noted that there was a considerable amount of overlap between this one microcluster and the other sparsely populated 'extra' microclusters. When the CluSandra framework's aggregator was run with an overlap factor of 1.0, that one microcluster absorbed all the extra sparsely populated microclusters. There was also a rather even distribution of data records across all five microclusters. A second series of experiments was executed that was identical to the first; the one exception being that the generator was configured to produce 7 classes, instead of 5. The results were consistent with those of the previous series of experiments.

VI. CONCLUSIONS AND FUTURE WORK

This work presents a distributed framework and algorithm for clustering evolving data streams. The Java-based framework, which is named the CluSandra framework, exhibits the following characteristics:

- It is entirely composed of proven open source components that can be deployed on a variety of commodity hardware and operating systems; therefore, it is very economical to implement.
- It is leveraged by clustering processes to address the severe time and space constraints presented by the data stream. In other words, the functionality required to address these constraints is offloaded from clustering process to the framework, and it is the framework that controls the data stream.
- It provides a distributed platform through which ensembles of clustering processes can be seamlessly distributed across many processors. This results in high levels of reliability, scalability, and availability for the clustering processes.
- It allows its hosted ensembles of clustering processes to elastically scale up and down to meet the most demanding dynamic data stream environments.
- It provides convenience classes or objects whose purpose is to facilitate the implementation of clustering algorithms and their subsequent deployment onto the framework.
- It provides an effective mechanism through which the hosted clustering processes can reliably and efficiently store their byproduct of real-time statistical summary information (i.e., microclusters) in a cluster database.
- It provides a Cluster Query Language (CQL) that is used to perform near-time or offline analytics against the cluster database. The CQL offloads the offline analytics from the clustering algorithm and provides a mechanism for the implementation of a variety of offline analytical processes. The CQL can also be used by offline processes to monitor, in near-time, clusters in the database and raise alerts whenever clusters of a particular nature appear and/or disappear.
- It lays down the foundation for a data management system whose focus is on clustering high speed and evolving data streams.

The algorithm developed is named the CluSandra algorithm and it is based upon concepts, structures, and algorithms introduced in [10] and [5]. The algorithm's implementation serves as an example of a clustering process that is designed to leverage the services and functionality provided by the CluSandra framework. The algorithm is more closely related to CluStream with its concept of viewing the data stream as a changing process over time and its functionality being divided between two operational phases: real-time statistical data collection and offline analytical processing. However, it deviates from CluStream primarily in how it addresses these operational phases. Unlike CluStream, it only performs the real-time statistical data collection, in the form of microclustering, leaving the offline analytical processing to end-users and/or processes that leverage the CQL. This results in a simpler and higher performing algorithm; primarily, because the agglomeration of microclusters is not performed by the algorithm. It also provides added flexibility to the end-user and/or offline analytical processes, because it affords the opportunity to analyze the collected data prior to agglomeration. Also, due to its reliance on the framework, the CluSandra algorithm can reliably and efficiently persist its microclusters to a cluster database; this is addressed in neither [10] nor [5]. The end result is the development of a clustering algorithm that exhibits these characteristics: configurable, distributable, elastically scalable, highly available and reliable, and simpler to implement.

The following are topics for future work:

- The implementation of clustering algorithms designed to address other data types and distributions, besides numerical and Gaussian, and their deployment onto the framework.
- The introduction of an integration framework, such as [18], that allows for the quick implementation of messaging design patterns meant to further control the data stream and facilitate the implementation of multi-staged data stream processing. For example, patterns to address data processing steps such as cleansing, standardization and transformation, and patterns used for routing data records based on their content.
- The implementation of a graphical user interface that leverages the CQL and provides a visual representation of the data in the cluster database.
- Additional research and development on algorithms that automate the calculation of an optimal MBT for the target data stream.

REFERENCES

[1] C. Walls, Spring In Action, 2nd ed., Greenwick, CT: Manning Publications Co., 2008.

[2] R. Pressman, Software Engineering A Practitioners Approach, 7th ed., New York: McGraw-Hill, 2009.

[3] P. Domingos, G. Hulten, "Mining high-speed data streams", in Knowledge Discovery and Data Mining, 2000, pp.71-80.

[4] P. Domingos, G. Hulten, L. Spencer "Mining time-changing data streams", in ACM KDD Conference, 2001, pp.97-106.

[5] C. Aggarwal, J. Han, J. Wang, P.S. Yu, "A framework for clustering evolving data streams", in Proceedings of the 29th VLDB Conference, Berlin, Germany, 2003.

[6] J. Gama, Knowledge Discovery From Data Streams. Boca Raton, FL: Chapman & Hall/CRC, 2010.

[7] J. Gama, M. Gaber, Learning From Data Streams, Processing Techniques in Sensor Data, Berlin-Hiedelberg: Springer-Verlag, 2007.

[8] B. Bifet, R. Kirkby, R., Data Stream Mining, University of Waikato, New Zealand: Centre for Open Software Innovation, 2009 .

[9] E. Hewitt, Cassandra, The Definitive Guide. Sebastopol, CA: O'Reilly Media, Inc., 2010.

[10] T. Zhang, R. Ramakrishnan, M. Livny, "BIRCH: an efficient data clustering method for very large databases", in ACM SIGMOD, 1996, pp. 103-114.

[11] G. Hebrail, Introduction to Data Stream Querying and Processing, International Workshop on Data Stream Processing and Management, Beijing: 2008.

[12] J. Han, M. Kamber, Data Mining Concepts and Techniques, 2nd ed., San Francisco, CA: Morgan Kaufmann Publishers, 2006.

[13] B. Everitt, S. Landau, M. Leese, Cluster Analysis, 4th ed., London, England: Arnold Publishers, 2001.

[14] B. Bifet, R. Kirkby, Massive Online Analysis Manual. University of Waikato, New Zealand: Centre for Open Software Innovation, 2009.

[15] "The Apache Cassandra Project." Available: http://cassandra.apache.org, [Accessed Sept. 2011].

[16] "Apache ActiveMQ." Available: http://activemq.apache.org, [Accessed Sept. 2011].

[17] T. White, Hadoop: The Definitive Guide, Sebastopol, CA: O'Reilly Media, Inc., 2010.

[18] "Apache Camel." Available: http://camel.apache.org, [Accessed Sept. 2011].

[19] C. Plant, C. Bohm, Novel Trends In Clustering, Technische Universitat, Munchen Germany, Ludwig Maximialns Universitat Munchen, Germany, 2008.

[20] L. Kuncheva, "Classifier ensembles for detecting concept change in streaming data: overview and perspectives", in Proceedings of the Second Workshop SUEMA, ECAI, Partas, Greece, July 2008, pp. 5-9.

[21] C. Aggarwal, J. Han, J. Wang, P.S. Yu, "A framework for projected clustering of high dimensional data streams", in Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004, pp. 852-863.

AUTHORS' PROFILES

Jose R. Fernandez received his MSc (2011) in Computer Science at the University of West Florida and his BSc (Eng., 1983) in Computer and Information Sciences at the University of Florida. He has over 25 years of commercial software engineering experience and has held senior architecture and management positions at NCR, AT&T, and BEA Systems. He has also been instrumental in starting several companies whose focus was on the development of Java Enterprise Edition (JEE) software products. He is currently a senior consultant with special research interests in machine learning and data mining.

Eman M. El-Sheikh is the Associate Dean for the College of Arts and Sciences and an Associate Professor of Computer Science at the University of West Florida. She received her PhD (2002) and MSc (1995) in Computer Science from Michigan State University and BSc (1992) in Computer Science from the American University in Cairo. Her research interests include artificial intelligence-based techniques and tools for education, including the development of intelligent tutoring systems and adaptive learning tools, agent-based architectures, knowledge-based systems, machine learning, and computer science education.

Clustering: Applied to Data Structuring and Retrieval

Ogechukwu N. Iloanusi
Department of Electronic Engineering
University of Nigeria, Nsukka
Enugu State, Nigeria

Charles C. Osuagwu
Department of Electronic Engineering
University of Nigeria, Nsukka
Enugu State, Nigeria

***Abstract*—Clustering is a very useful scheme for data structuring and retrieval behuhcause it can handle large volumes of multi-dimensional data and employs a very fast algorithm. Other forms of data structuring techniques include hashing and binary tree structures. However, clustering has the advantage of employing little computational storage requirements and a fast speed algorithm. In this paper, clustering, k-means clustering and the approaches to effective clustering are extensively discussed. Clustering was employed as a data grouping and retrieval strategy in the filtering of fingerprints in the Fingerprint Verification Competition 2000 database 4(a). An average penetration of 7.41% obtained from the experiment shows clearly that the clustering scheme is an effective retrieval strategy for the filtering of fingerprints.**

Keywords-component; Clustering; k-means; data retrieval; indexing.

I. INTRODUCTION

A collection of datasets may be too large to handle and work on hence may be better grouped according to some data structure. Large datasets are encountered in filing systems in digital libraries, access to and caching of data in databases and search engines. Given the high volume of data there is need for fast access and retrieval of required or relevant data. Several of the existing data structures are hashing [1, 2, 3, 4, 5, 6], search trees [7, 8], and clustering [9]. Hashing is a technique that utilizes a hash function to convert large values into hash values and maps similar large values to the same hash values or keys in a hash table. Clustering is however a useful and efficient data structuring technique because it can handle datasets that are very large and at the same time n-dimensional (more than 2 dimensions) and similar datasets are assigned to the same clusters [9]. A 2D or 3D point can be imagined and illustrated however it will be difficult to imagine or illustrate a 9-dimensional data. When datasets are clustered, the clusters can be used rather than the individual datasets.

Clustering is a process of organizing a collection of data into groups whose members are similar in some way [9, 10, 11, 12] According to Jain et al. [13] "Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity". Similarity is determined using a distance measure and objects are assigned and belong to the same cluster if they are similar according to some defined distance measure. Cluster analysis differs from classification because in clustering the data are not labeled and hence are naturally partitioned by the clustering algorithm whereas in classification the data are labeled and partitioned according to their labels. The former is hence an unsupervised mode of data structuring while the later is supervised [13]. Jain [14] identifies three main reasons while data clustering is used; to understand the underlying structure of the data; to determine degree of similarity amongst the data in their natural groupings and to compress data by summarizing the data by cluster groups.

Clustering has a vast application in the life sciences, physical and social sciences and especially in the disciplines of Engineering and Computer Science. Clustering is used for pattern analysis, recognition and classification, data mining and decision making in areas such as document retrieval, image processing and statistical analysis and modeling [13]. Documents may be clustered for fast information access [15] or retrieval [16]. Clustering is used in image processing to segment images [17] as well as in marketing, biology, psychiatry, geology, geography and archeology [13]. Figure 1 shows a general data clustering illustration. The data are grouped in clusters. Each cluster has a collection of data that are similar.

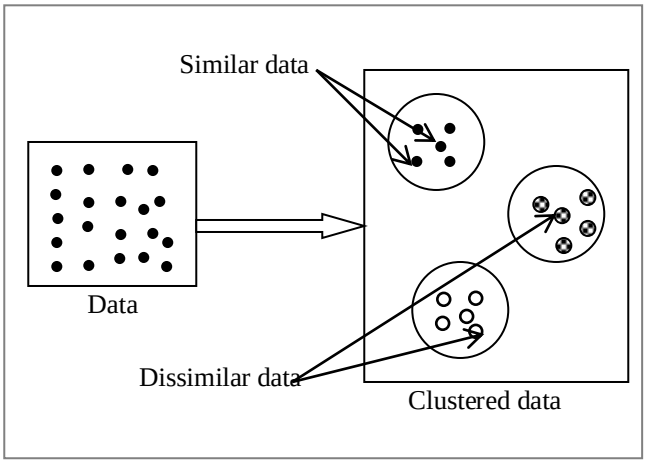


Figure 1. Data Clustering

A cluster is a group of similar datasets represented by an n-dimensional value given by the cluster centroid. Clusters may also be defined as "high density regions separated by low density regions in the feature space" [13].

Every cluster is assumed to have a centroid, which is the arithmetic mean of all data in that cluster. The mean is what is common to data assigned to a cluster and creation of clusters

builds from the arithmetic mean. A similarity measure is used for the assignment of patterns or features to clusters.

II. Cluster Similarity Measures

Similarity is fundamental to the definition of a cluster hence a measure for the similarity otherwise known as the distance measure is essential. The dissimilarity or similarity between points in the feature space is commonly calculated in cluster analysis [13]. Some of the distance measures used are:

- Euclidean distance
- Manhattan distance
- Chebyshev distance
- Hamming distance

The distance metric is used for computing the distance between two points and cluster centers. For the distance measures explained in the following sections, two points, a and b, are defined in an n-dimensional space as:

$$\mathbf{a = (w_0, x_0, y_0 \ldots z_0)\ coordinates} \quad (1)$$

$$\mathbf{b = (w_1, x_1, y_1 \ldots z_1)\ coordinates} \quad (2)$$

A. Euclidean distance

Euclidean distance is the distance between two points, a and b, as the crow flies in an n-dimensional space.

$$\boldsymbol{D(a,b) = \sqrt{((w_0 - w_1)^2 + \cdots + (z_0 - z_1)^2)}} \quad (3)$$

$$= \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (4)$$

where n is the number of dimensions. The Euclidean distance is the most commonly used metric because it is appealing to use in an n-dimensional space and it works well with isolated clusters [13].

B. Manhattan distance

In the Manhattan distance, the distance between two points is the absolute difference of their coordinates.

$$\boldsymbol{D(a,b) = (w_0 - w_1) + \cdots + (z_0 - z_1)} \quad (5)$$

The difference between the Euclidean distance and the Manhattan distance is that the Euclidean is a squared distance while the Manhattan is not squared.

C. Chebyshev distance

In the Chebyshev distance metric the distance between two points is the greatest of their differences along any coordinate dimension [18]. This distance is named after Pafnuty Chebyshev.

$$\boldsymbol{D(a,b) = \max(a_i - b_i)} \quad (6)$$

This is also known as the chessboard distance. In the chessboard the length of side of a chess square may be assumed as one unit. In this case the minimum number of moves needed by a king to go from one chess square to another equals the Chebyshev distance between the centers of the squares.

D. Hamming distance

The Hamming distance is a way of determining the similarity of two strings of digits of equal lengths by measuring the number of substitutions required to change a string into another. It is the number of positions at which corresponding digits in the two strings are different [19].

Given two strings a and b where

a = 0110110 and b = 1110011, the difference between the two strings a and b, D(a,b), where

$$\begin{array}{l} a = 0\,1\,1\,0\,1\,1\,0 \\ b = 1\,1\,1\,0\,0\,1\,1 \end{array}$$

D(a,b) = 3, as the corresponding digits differ in three places.

III. CLASSIFICATION OF CLUSTERING ALGORITHMS

Clustering algorithms may be classified as:

- Exclusive clustering
- Overlapping clustering
- Hierarchical clustering

A. Exclusive clustering

In exclusive clustering, data that belongs to a particular cluster cannot belong to another cluster. An example is K-means clustering.

B. Overlapping clustering

Data may belong to two or more clusters. Example of this in fuzzy-c-means clustering.

C. Hierarchical clustering

In this case clusters are represented in tree from. Two close clusters are derived from the top-level cluster. The hierarchy is built by individual elements progressively merging into bigger clusters.

Figure 2 shows the types of data clustering algorithms.

Jain [13] classifies clustering algorithms as hierarchical and partitional. In hierarchical clustering each cluster arises from and depends on the parent cluster. A typical partitional clustering algorithm is the K-means algorithm.

IV. Cluster Similarity Measures

K-means clustering algorithm was first proposed over 50 years ago [14] and is commonly preferred to other clustering algorithms because of its ease of implementation and efficiency in cluster analysis.

K-means clustering is a type of cluster analysis that partitions n observations into k disjoint clusters, k<<n, such that the number of clusters are much less than the number of observations [18, 20]. The k-means algorithm partitions n observations $\{O_i \mid i=1, 2...n\}$ into k number of clusters, $\{C_j \mid j=1,2...k\}$, as follows

$$\{\boldsymbol{C_j \in O_i}\}\ \boldsymbol{if}\ \{\boldsymbol{argmin} \sum_{j=1}^{k} \sum \left\| \boldsymbol{O_i - C_j} \right\|^2\} \quad (7)$$

This is illustrated in Figure 3.

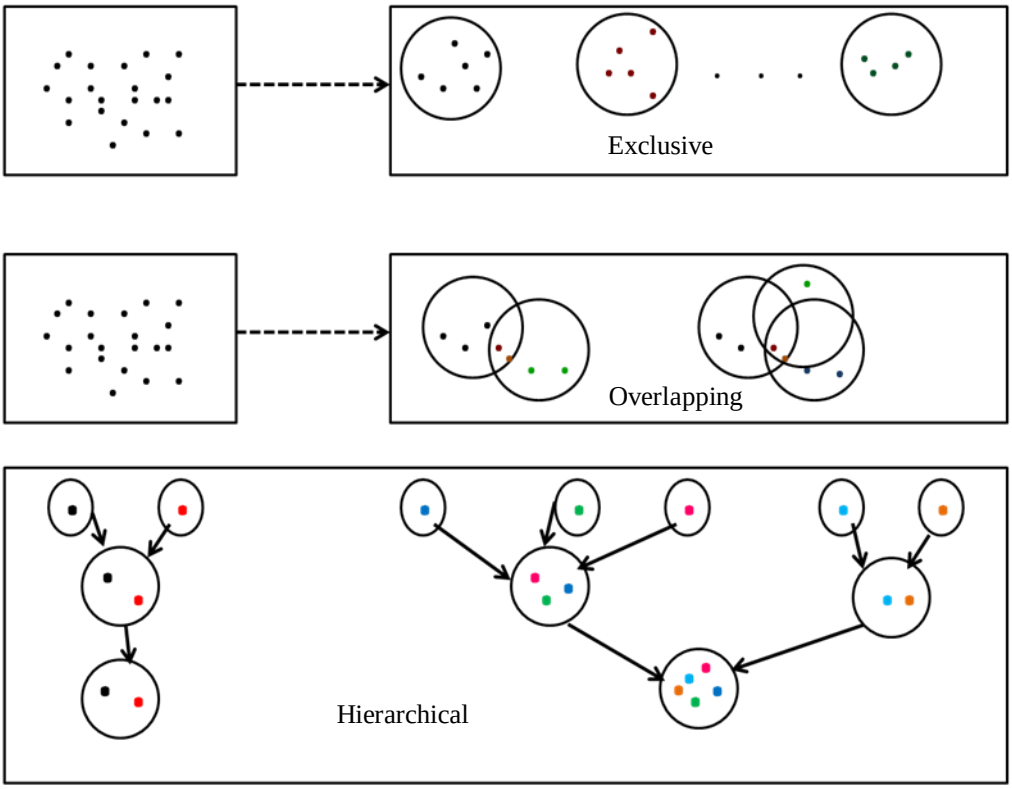


Figure 2. Types of Data Clustering Algorithms

Any of the distance metrics which include the Euclidean, Manhattan, Chebyshev or Hamming may be used as the distance measure for determining the similarity of the datasets, though the Euclidean is most preferred and widely used [14].

The K-means algorithm basically follows these steps.

- A similarity or distance measure is chosen and used throughout.
- K number of centroids are chosen.
- The distance between each dataset from each of the k centroids is determined.
- Then a dataset is assigned to the centroid for which it had the minimum distance.
- All datasets are hence assigned to a particular centroid. Figure 4 shows a very simple illustration using 5 datasets and 2 clusters.
- The arithmetic mean is recalculated for each of the k centroids and the distance of each dataset from the new means is recalculated for each of the k centroids. This is the second iteration.
- The datasets are reassigned again to the new k centroids. In other words, a dataset assigned for instance to centroid 2 in the first iteration may be reassigned to centroid 1 in the second iteration.

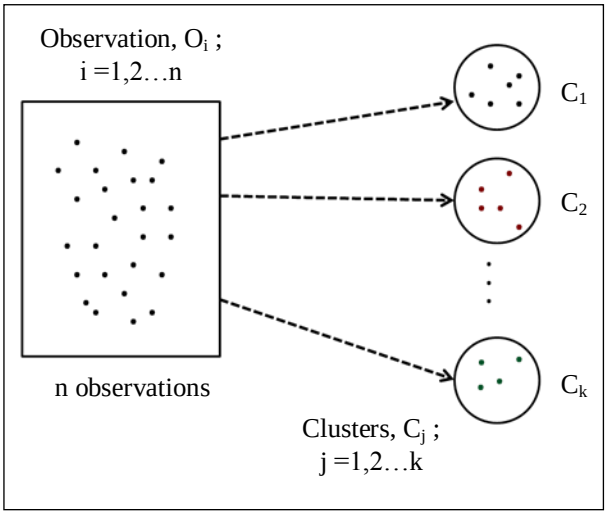
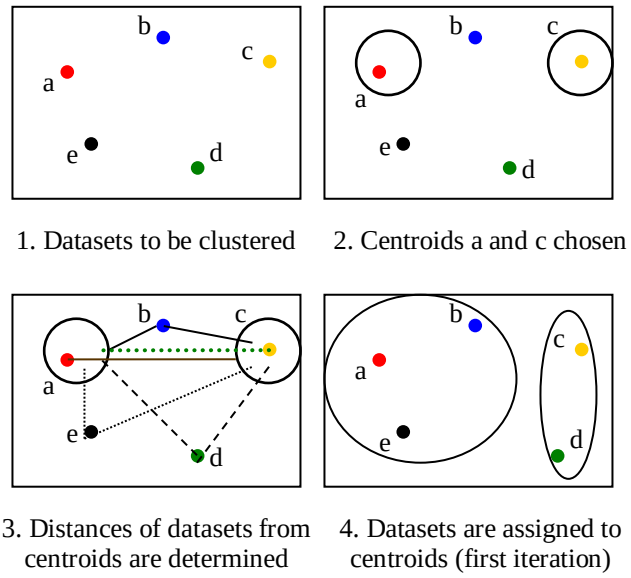


Figure 3. K-means Clustering

- The arithmetic means is recalculated and the data-sets reassigned again.
- This continues to i number of iterations, and the iteration stops when there is no change in the assignments between the ith iteration and the (i-1)th iteration.
- The last k centroids are the k clusters.



1. Datasets to be clustered
2. Centroids a and c chosen
3. Distances of datasets from centroids are determined
4. Datasets are assigned to centroids (first iteration)

Figure 4. K-means clustering algorithm

V. Approaches to Effective Cluster Analysis

Several challenges to cluster analysis include the difficulties in choosing an appropriate clustering algorithm; representing the data to be clustered; choosing a suitable similarity measure; determining which data should be used and choosing a suitable number of clusters that would yield maximum success. A user is not faced with these problems in hierarchical clustering analysis since all the datasets are related. These problems arise when a dataset is to be classified into unique clusters.

There are many partitional clustering algorithms and the user may be faced with the dilemma of choosing an appropriate algorithm. What guides in choosing an appropriate algorithm is to know the purpose and goal of the clustering exercise and this consequently would guide in representing that data to be clustered. It is also necessary to know if the dataset has a clustering tendency [18] and if it should be normalized. A data set that does not have a clustering tendency should not be clustered as it would yield invalid clusters. For example, if a data set that has all similar data and consequently has no variance is clustered it would result in invalid clusters. On the other hand a data set with high variance has a clustering tendency.

The choice of number of clusters may pose a problem because the performance of the clustering algorithm is affected by the number of clusters. It is usually difficult to determine the best number of partitions that will give the best and valid clustered groups. Some factors that may be considered while choosing the number of clusters are the size of the data set and the variance of the data in the data set. If the dataset is widely varied such that the data set may need to be classified by many groups then it may make sense to use more clusters.

In feature classification, the success of the cluster analysis is largely dependent on the feature set. The clustering algorithm would have a good performance and give compact, isolated and valid clusters if the choice of features is good [18]. If for instance a database of face images of a multi-racial group comprising African, Chinese, Latin American, Indian and European faces need to be clustered into five different groups, the features that would be used would be such that the faces can be effectively separated into five valid clusters. The success of this task is clearly dependent on the features used for the separation. The features making up the data set play a vital role in clustering analysis.

A similarity measure is required for separating data into clusters. The choice of the similarity measure is a challenging problem because the valid clustering of the data also depends on the similarity metric. The performance of the cluster analysis varies according to the similarity metric used and hence it may be difficult to determine the similarity metric that would give the best performance. But this problem can be overcome by having a good understanding of the data to be clustered.

VI. Clustering Used as a Fingerprint Indexing Retrieval Strategy

An indexing technique must include a retrieval strategy. A retrieval strategy defines the method for which data within the same class as the query or input data are retrieved. In fingerprint indexing, the retrieval strategy ensures that fingerprints with similar index codes [21] to that of the query fingerprint are retrieved from the database of enrolled fingerprints.

In this work, a modified Ross's partitional clustering scheme [22] is used as a fingerprint retrieval strategy by compressing the numerous fingerprint features into similar groups of data and hence limiting the search for similar fingerprints to only a few clusters that are identical to the cluster of the query fingerprint. This requires the following

- First the creation of an index space of k clusters for the indexing using the k-means algorithm and the Euclidean distance similarity measure.
- Secondly the assignment of the features of the fingerprints in the enrolled database to the k clusters.
- Thirdly the determination of the clusters, c << k, that have the features of the fingerprints similar to a query fingerprint.

A query fingerprint should have a matching identity in a list of fingerprints outputted by the indexing algorithm. This list is otherwise known as the candidate list. The ratio of the fingerprints in the candidate list to the database size gives the penetration rate of a query fingerprint. The penetration rate is the fraction of fingerprint identities, including the genuine fingerprint, retrieved from the database upon presentation of an input fingerprint. The penetration rate determined for a number of tests, T, in a database of size, N, is [23].

$$\textbf{penetration rate} = \frac{1}{T}\sum_{j=1}^{T}\frac{C_j}{N} \qquad (8)$$

Where $\{C_j \mid j = 1, 2 \ldots T\}$ is the size of candidate list of the fingerprints. The less the penetration rate the better the performance of the algorithm.

In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use

VII. Experiments and Results

- The Fingerprint Verification Competition (FVC) 2000 database 4(a) and FVC 2002 database 4(a) were used for this experiment. Each database has 800 fingerprints from 100 subjects at 8 impressions per subject.
- The fingerprint features were extracted using the minutiae quadruplets technique [24].
- 30 clusters were created in the index space using fingerprints from FVC 2002 database 4(a).
- The FVC 2000 database 4(a) was divided into two equal groups – Group A and B.
- Group A had 400 fingerprints of the first four impressions of a subject
- Group B had 400 fingerprints of the last four impressions of a subject.
- The fingerprint features of group A were assigned to the 30 clusters in the index space.
- The fingerprints of group B were used to query the index space to find a matching identity determined by the penetration of the database. Every query resulted in a penetration rate. Majority of the queries had little penetration rates while some had long penetration rates.

The penetration rates of the 400 query fingerprints used in the experiments are shown in Table 1.

TABLE I. Penetration rates of 400 query fingerprints in an evaluation on FVC 2000

No of tests (T)	Penetration rates (range)	Average value	fx
57	0.25 - 1.00	0.75	42.75
55	1.25 - 2.00	1.625	89.375
39	2.25 - 3.00	2.625	102.375
49	3.25 - 4.50	3.875	189.875
50	4.75 - 6.50	5.625	281.25
47	6.75 - 8.50	7.625	358.375
55	8.75 - 14.25	11.5	632.5
48	14.5 - 38.25	26.375	1266
T=400			**∑fx=2962.5**

The average penetration for the 400 query fingerprints is obtained using Equation (8) and can also be determined from Table 1 as:

$$Average\ penetration = \frac{\sum f_x}{T} = \frac{2962.5}{400} = 7.406\% \quad (9)$$

Where f_x is the product of the first and third columns in Table I and T is the number of queries corresponding to the number of tests in the experiment. There were 400 queries.

The retrieval of a candidate list for a query fingerprint takes 0.592ms.

VIII. Comparison with Other Data Structuring Techniques

In [25], a binary tree based approach was used for matching fingerprints. The work done on this paper is indexing. However, the computational time for a fingerprint match using the binary tree technique in [25] is compared with the computational time for indexing a query fingerprint using the clustering technique described in this paper in Table II.

TABLE II. Comparison of computational times of the binary tree and clustering techniques on FVC 2002 DB1 SET A

Technique	Database size	Computational time
Binary tree [25]	800 fingerprints	34.8ms
Clustering (our approach)	400 fingerprints	0.592ms

IX. Conclusion

In this paper, clustering was discussed extensively. Experiments were conducted by employing a modified clustering scheme as a retrieval strategy for filtering fingerprints. The average penetration, 7.41%, is very small showing clearly that the clustering algorithm employed is an effective scheme for the filtering and retrieval of the candidate fingerprints to a given query fingerprint.

References

[1] H. J. Wolfson, "Geometric Hashing: An Overview" IEEE Computational Science and Engineering, pp. 10 – 21, 1997.

[2] S. Danker, R. Ayers and R. P. Mislan, "Hashing Techniques for Mobile Device Forensics" SMALL SCALE DIGITAL DEVICE FORENSICS JOURNAL, vol. 3, no. 1, June 2009 ISSN:1941-6164

[3] D. Zhang and J. Wang, "Self-Taught Hashing for Fast Similarity Search" SIGIR'10, July 19–23, 2010, Geneva, Switzerland

[4] H. Lim, J. Seo and Y. Jung, "High speed IP address lookup architecture using hashing" IEEE Communication Letters, vol. 7, no. 10, pp. 502 – 504, ISSN: 1089-7798, October 2003

[5] D. Greene, M. Parnas and F. Yao, "Multi-index hashing for information retrieval" Proceedings of the 35th Annual Symposium on the Foundations of Computer Science, 1994, pp. 722 – 731

[6] X. Nie, J. Liu, J. Sun and W. Liu, "Robust Video Hashing Based on Double-Layer Embedding" IEEE Signal Processing Letters, vol. 18, no. 5, pp. 307 – 310, May 2011

[7] B. Pfaff, "Performance Analysis of BSTs in System Software" SIGMETRICS/Performance'04, New York, NY, USA. June 12–16, 2004,

[8] D. D. Sleator and R. E. Tarjan, "Self-Adjusting Binary Search Trees" Journal of the Association for Computing Machinery. vol. 32, no. 3, July 1985, pp. 652-686.

[9] A.K. Baughman, S. Stockt, and A. Greenland, "Large Scale Fingerprint Mining." 2010 ACM 978-1-4503-0220-3.

[10] A. K. Jain, A. Topchy, M. Law, and J. Buhmann, "Landscape of Clustering Algorithms", in Proceedings of the 17th International Conference on Pattern Recognition, Cambridge UK, August 23-26, 2004 pp. I-260--I-263.

[11] F. Alberto and G. Sergio, "Solving Non-uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms". Journal of Classification, Vol.25, 2008, pp.43–65.

[12] D. Apetrei, P. Postolache, N. Golovanov, M. Albu and G. Chicco, "Hierarchical Cluster Classification of Half Cycle Measurements in Low Voltage Distribution Networks for Events Discrimination." International Conference on Renewable Energies and Power Quality (ICREPQ'09).

[13] A. Jain, M. N. Murty and P. Flynn, "Data clustering: A review." ACM Computing Surveys, Vol.3, no. 13, 1999, pp. 264–323.

[14] A.K. Jain, Data Clustering: "50 Years Beyond K-Means." Pattern Recognition Letters, Vol. 31, No. 8, 2010, pp. 651-666.

[15] M. Sahami, "Using Machine Learning to Improve Information Access." Ph.D. Thesis, Computer Science Department, Stanford University. 1998.

[16] S. Bhatia and J. Deogun, "Conceputal clustering in information retrieval." IEEE Transactions. Systems Man Cybernet. Vol. 28 (B), 1998, pp. 427–436.

[17] A. K. Jain and P. Flynn, "Image segmentation using clustering." in Advances in Image Understanding. IEEE Computer Society Press, 1996, pp. 65–83.

[18] R.M.C.R. Souza and F.A.T. Carvalho, "Dynamic clustering of interval data based on adaptive Chebyshev distances." Electronics Letters, Vol 40, Issue 11, ISSN: 0013-5194, 2004, pp. 658 – 660.

[19] L. Gąsieniec, J. Jansson and A. Lingas, "Approximation algorithms for Hamming clustering problems." ELSEVIER. Journal of Discrete Algorithms, Vol. 2, 2004, pp. 289–301.

[20] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, July 2002. pp 881-892.

[21] D. Maltoni, D. Mario, A.K. Jain, and S. Prabhakar, "Handbook of Fingerprint Recognition." 2nd Ed. Springer-Verlag London Limited. 2009.

[22] A. Ross and R. Mukherjee, "Augmenting Ridge Curves with Minutiae Triplets for Fingerprint Indexing." In: Proc. of SPIE Conference on Biometric Technology for Human Identification IV. Orlando, USA. (April 2007).

[23] A. Gyaourova and A. Ross, "A Novel Coding Scheme for Indexing Fingerprint Patterns." In Proc. Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition. Springer-Verlag, 2008, pp 755 -764.

[24] O. Iloanusi, A. Gyaourova, A. Ross, "Indexing Fingerprints Using Minutiae Quadruplets," Proc. of IEEE Computer Society Workshop on Biometrics at the Computer Vision and Pattern Recognition (CVPR) Conference, (Colorado Springs, USA), 20-25 June 2011, pp 127 - 133.

[25] M. D Jain, N. S. Pradeep, C. Prakash and B. Raman, Binary tree based linear time fingerprint matching, IEEE International Conference on Image Processing (ICIP), 2006

AUTHORS PROFILE

Dr. Ogechukwu Iloanusi is a Lecturer at the Department of Electronic Engineering, University of Nigeria, Nsukka. She holds a Ph.D in Digital Electronics and Computer. Her research interests are biometric recognition and algorithm development, web-based applications, micro-processor based system design and e-learning..

Prof. Charles Osuagwu is a Professor of the Department of Electronic Engineering, University of Nigeria, Nsukka. He has an M. Sc and Ph.D in Engineering from the University of Southampton. His major research interests are Digital Signal Processing, Microprocessor-based System Design, Software Design, Computer Design, e-Security and Institutional Improvement.

Irrigation Fuzzy Controller Reduce Tomato Cracking

Federico Hahn
Irrigation Department
Universidad Autonoma Chapingo
Chapingo, Texcoco, México

***Abstract*—Sunlight heats the greenhouse air temperature and tomato cracking decreases marketable product up to 90%. A shade screen reduced incoming radiation during warm and sunny conditions to reduce tomato cracking. A fuzzy controller managed greenhouse irrigation to reduce tomato cracking using as variables solar radiation and substrate temperature. The embedded controller presented 9 rules and three assumptions that made it operate better. Signal peaks were removed and control actions could take place ten minutes after irrigation. Irrigation was increased during the peak hours from 12:00 to 15:00 h when it was required by the fuzzy controller; meanwhile water containing the nutrient solution was removed during very cloudy days with limited photosynthesis. After three continuous cloudy days irrigation should be scheduled to avoid plant nutrient problems. Substrate temperature in volcanic rock can be used as real time irrigation sensor. Tomato cracking decreased to 29% using the fuzzy controller and canopy temperature never exceeded 30ºC.**

Keywords-component; Fuzzy controller; irrigation controller; tomato cracking.

I. INTRODUCTION

Sunlight intensity control decrease irrigation needs as plants and soil are kept cooler, increasing plant growth as temperature drops off [1]. Cuticle cracking is a physiological disorder that occurs at the beginning of the last phase of fruit growth decreasing skin elasticity [2]. Tomato becomes unmarketable for fresh consumption with losses ranging from 10% to 95% of the total fruit. Summer crack unmarketable tomatoes are caused by high air temperature within the greenhouse [3], so shading becomes an alternative, although tomato yield decreases. Retractable roof shade houses are structures covered with polypropylene, polyethylene, or composite fabrics save up to 30% in energy [4]. Screen systems reduce heat radiation losses at night, and decrease the energy load on the greenhouse crop during sunny conditions. Constant daily light integral achieved only through shading or supplemental lighting [5], induces plant growth passing from its initial transplanting state to its harvest state in 25 days.

Most of the irrigation systems uses ON/OFF controllers, but optimal results are difficult to achieve due to varying time delays and system parameters. A computer-based system controlled irrigation and fertilization in greenhouses; sensors provided feedback information for operating pumps and solenoid valves [6]. Fuzzy theory interprets real uncertainties and becomes ideal for nonlinear, time varying and hysteretic system control. A fuzzy controller system saved water in greenhouses and was cheap to implement [7]. A greenhouse fuzzy climate controller required of 81 rules for proper operation [8]. A simple fertigation fuzzy control presented potential to save water and nutrients [9].

The following paper describes the design, implementation of a fuzzy greenhouse controller which monitored radiation and substrate temperature. Additional nutrients and water were supplied to the greenhouse by an irrigation control to reduce tomato cracking during the hot summer.

II. MATERIALS AND METHODS

The 80 x 40 m University greenhouse at Tlapeaxco, Chapingo was used for the experiment, having natural cooling through lateral and cenital openings. A black Raschel net having a 50% of transmission was opened manually during the first three months and closed during the summer when tomatoes were produced. Three rows along the greenhouse were transplanted on March 2th, 2009 in volcanic rock substrate (dimensions: 40 x 40 x 10 cm) with tomato seedlings (*Lycopersicon esculentum* var. Roma). A solenoid controlled water and nutrient application in each row having tomato plants spaced 30 cm; the irrigation treatments were: A, B and control. Treatment A used the fuzzy controller to apply additional irrigation during hot sunny days and removed water under cloudy conditions. Treatment B added water with the fuzzy controller under whichever weather condition. The last treatment used the conventional controller and did not use any kind of shading. Red tomatoes were harvested on the first of June and one month later cracking was evaluated per treatment.

A fuzzy controller was selected for this application as precise mathematical modeling of the controlled object is not required becoming simpler to implement. Two photometric sensors (model LI-210SB, Li-Cor Environmental Division, USA) measured illumination providing 10 mV/100 klux being cosine corrected up to 80 degree angle of solar incidence. The sensors were calibrated against a standard lamp and presented a sensitivity of 20 uA per 100 klux and a response time of 10 microseconds. One sensor was placed in the shading area, and the other in the area without control. Radiation was acquired every thirty seconds averaging ten continuous measurements.

Substrate temperature was measured with one thermocouple fixed horizontally 8 cm from the pot top and beneath the dripper. The J cooper constantan type thermocouple was connected to a module (model TxRail 4-20 mA, Novus, Brasil) to provide a 4-20 mA signal finally converted by means of a resistance to voltage. Platinum RTD sensors (model RP502T22, Advanced Thermal Products, USA) measured canopy temperature and was fixed beneath the

leaves. Six substrate and canopy temperature measurements were acquired every 30 seconds from six plants selected randomly. Temperature and radiation values were stored in a datalogger (model CR1000, Campbell Scientific, USA). Daily stored data were analyzed, comparing treatments behavior and when high differences were noted a revision on the sensor network and irrigation drippers was carried out.

A. Fuzzy membership functions and rules

Fuzzy evaluation methods process all the variables according to predetermined weights and decrease the fuzziness by using membership functions; therefore sensitivity is quite high compared to other index evaluation techniques. The fuzzy input membership functions determine the variables that are going to be acquired to develop the control. It is desirable to reduce the number of functions in order to minimize the number of rules and the data contained in microcontroller memory. Two input membership functions were used; the first one employed the substrate temperature and the second one the radiation, "Fig. 1". The temperature function presented three groups: COOL, FRESH and WARM temperature, being the latter trapezoidal for values over 20°C. The second function is composed of three radiation groups named sunny (SUN), cloudy (CLOU) and very cloudy (VC). The very cloudy group of black and rainy sky presents the lowest radiation; a clean sky radiates strongly and forms part of the sunny triangle.

Once the variables are acquired several rules establish the actions that control the greenhouse tomato crop, Table I. The two control actions are to indicate when to irrigate and to increase irrigation and nutrient application. For example a WARM temperature AND sunny radiation will irrigate the crop (IRR) and add an additional 30% of water and 15% of nutrients (ADNUT). In the case of a COOL temperature AND very cloudy radiation (VC) the plants will not be irrigated (NON IRR) and zero additional nutrition applied (ZNUT) due to poor water and nutrient uptake. The AND condition was done with the minimum algorithm. A real time clock controlled the fuzzy rules during the 9:00-17:00 hr period. Every one and a half hour starting at 9 AM drip irrigation was applied for seven minutes. A total of seven periods were programmed per day, but the fuzzy controller could apply five minutes more during the irrigation periods of 12:00, 13:30 and 15:00, increasing water application by 30%.

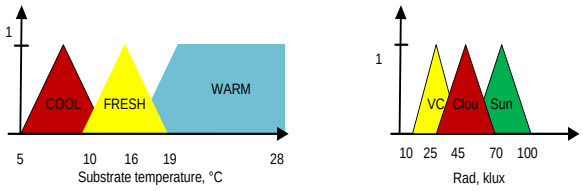


Figure 1. Temperature and radiation membership function

An embedded system used an ATM89C51 micro-controller to control the solenoids to regulate water application per row.

TABLE I. FUZZY RULES TO CONTROL THE IRRIGATION SYSTEM

Radiation	Nutrient and Water application		
	Cool	*Fresh*	*Warm*
VC	ZNUT/NON IRR	ZNUT/NON IRR	ZNUT/NON IRR
SC	ZNUT/ NON IRR	ADNUT/ IRR	ZNUT/ IRR
S	ZNUT/ IRR	ADNUT/ IRR	ADNUT/ IRR

Substrate temperature values between 5 and 28°C were converted to a signal varying from 0 to 5 volts. The ADC804 converter acquired the temperature and radiation signals and the digital output was sent to port 1. The real time clock (DS 1706) was connected to port 3. The fuzzy values were transformed to its membership function using truth tables. Irrigation control turned-on the pump and the solenoids, according to the fuzzy rules used for adding water and nutrients. The same fuzzy rules were carried out by treatments A and B, but the microcontroller did not remove irrigation on days with limited illumination in treatment B.

III. RESULTS AND DISCUSSION

Soil temperature inside the greenhouse nearby the substrate was measured at 9:00, 12:00 and 15:00 h, Table II. At noon the soil temperature covered with the black screen was 5.1°C lower than soil exposed directly to the sun. Air temperature measured at a height of 3 m was 7.2°C hotter than the soil temperature. Inside greenhouses, leaves can be cooler than air decreasing transpiration and increasing leaf condensation.

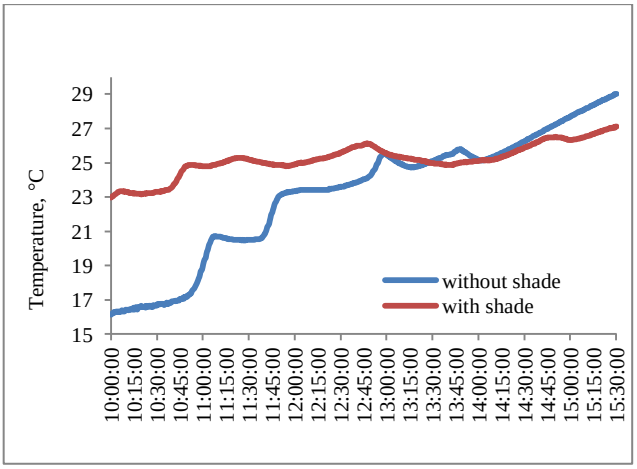


Figure 2. Substrate temperature affected by the shade screen.

Substrate temperature with and without shade becomes alike at 1:00 PM; during the morning shaded substrate temperature is higher, "Fig. 2". Blue line discontinuities result from irrigation. Temperature difference between canopy and substrate was higher when no shading was present and maximum values were encountered in sunny mornings. Many peaks were encountered during unsteady climatic conditions, causing multiple screen movement. Three assumptions were considered for a better controller performance:

1. At least three peaks were required in the same direction during a period of 15 minutes
2. If one peak occurred in one direction and the next in opposite direction they were cancelled.
3. After irrigation a control action could take place after 10 minutes.

TABLE II. SOIL TEMPERATURE MEASURED AT THREE DIFFERENT HOURS IN A BLACK SHADED GREENHOUSE.

	Soil temperature, °C		Light, klux	Air temp, °C
	Full sun	*30% Black shade*		
09:00 a.m.	17.6	19.5	16	28
12:00 a.m.	38.2	33.1	70	40.3
03:00 p.m.	42	37.7	62	42

The three conditions used by the fuzzy controller under operation made it perform better removing instantaneous temperature and radiation peaks. The third rule was the most useful as substrate temperature changes occurred just after irrigation, affecting the substrate temperature signal.

The fuzzy controller was tested when step irrigation was applied; canopy temperature changed after fifteen minutes, "Fig. 3". Substrate temperature decreased by 5°C after the twelve minute irrigation session (7 minutes of the normal cycle + 5 minutes of the control) in the cloudy afternoon. Although water is applied for twelve minutes, substrate temperature variations will be noted during 20 minutes. Substrate temperature difference between cloudy and sunny days recorded every 5 min and averaged over each 60-min interval were encountered at the moment of maximum temperature [10]. The fuzzy controller takes ten data and after processing one value will appear every 5 minutes. Substrate daily temperature surpassed soil temperature and when it exceeded 35°C for more than 6 hours/day plant injury was expected decreasing internodes spacing [11].

Maximum substrate temperature achieved throughout the month was of 30°C and reduced with irrigation as noted in Fig. 3. If root water absorption is less than the rate of transpiration loss, stomata will close, increasing plant and air temperature; transpiration rate reduction in a shaded greenhouse decreased total crop water uptake by 33% [12].

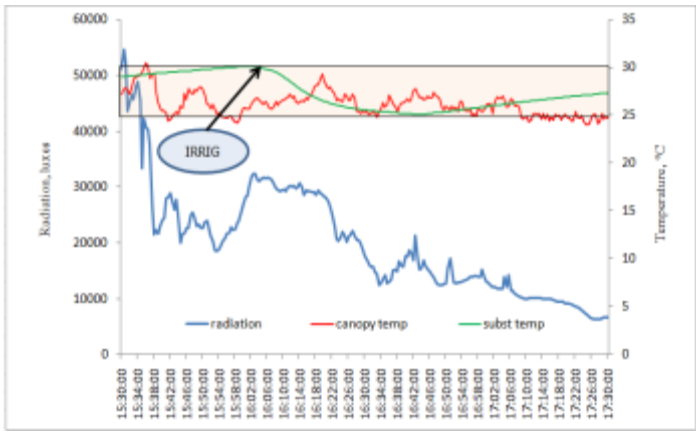


Figure 3. Radiation and temperature differences measured after irrigation.

Substrate temperature was considered a good indicator of canopy temperature during all the day except during irrigation. It can be used as a real time irrigation sensor for volcanic rock substrate characterized by air spaces and low water retention. Thirty hours were not irrigated due to electricity problems, increasing the differential temperature between canopy and substrate to 24°C due to very little transpiration; Screens were shot reducing the incoming radiation to the plants. Application of continuous water in dark rainy days increased moisture to 80-90% so less irrigation was required and nutrients were applied only twice a day. After three continuous cloudy days irrigation should be scheduled for treatment A to avoid plant nutrient problems.

Cuticle cracking appears under high illumination conditions when temperature increases inside the greenhouse. Daily tomato growth variations and tomato heating caused by direct radiation cracked 52% of the tomatoes as noted on fruits screen-uncovered; rapid tomato daily growth is avoided and can withstand environmental and nutritional variations during tomato growth. Cracking was reduced from 52% obtained during the harvest of the first of June to 29% at the harvest of the first of July using the fuzzy controller with treatment A. If water lost by transpiration could not be replenished by the root system, it will use water stored inside the fruit [13]. During night, lower leaf transpiration and root pressure would resupply water to the fruit. Factors such as volume and irrigation frequency affect water uptake by the plants and could be involved in the development of cracking [14]. Canopy temperature never exceeded 30°C as additional irrigation was added by the controller. Shading reduced the radiation maintaining fresh the fruits; cracked harvested fruit resulted from tomatoes having cracks at earlier stages.

Fruit weight increased when more irrigation was applied, but in sunny days with high transpiration, water arrival to the fruit decreased. Lack of irrigation once tomato weight exceeds 75 grams, affects its internal pressure and will start cracking the fruits under direct heating. Average weight of the tomato fruits with fuzzy control was 71 g being 6.4 g smaller than tomatoes grown without control. Weights of tomatoes grown under fuzzy control were more homogeneous as noted by the standard deviation values of 0.77 and 0.69 for both treatments. In treatment B, the fuzzy controlled supplied additional irrigation and maintained the irrigation cycles during limited illumination. As nutrient uptake by the plants is limited during scarce illumination electrical conductivity increased resulting in 31% of tomato cracking. The producer noted nutrient and water losses during treatment B compared with treatment A presenting both similar tomato average size. Plants without shade presented losses of 52% by cracking, and although their size was bigger and the yield higher. Different doses of irrigation were applied when half of the days were sunny and half cloudy being screen shading essential to reduce fruit cracking.

Future work includes a movable controlled shade screen as it was opened manually during this experiment; the screen should be opened or closed automatically with a gear motor driven by a photovoltaic system [15]. A pH and electrical conductivity (EC) fuzzy system is required to control

fertigation more precisely, helping to observe cracking caused by EC variations [9].

IV. CONCLUSION

This work concludes that tomato cracking can be reduced by setting a screen; Cracking tomatoes harvested during July were reduced from 52% to 29%. Raschel shading screens reduced incoming radiation, fruit size and yield, but increased marketable production and producer profit. The use of the fuzzy controller regulated canopy temperature which never exceeded 30°C. Irrigation applied on sunny days helped to maintain a constant fruit growth as well as a better canopy temperature. Less water and fertilizers were used in treatment A than in treatment B where irrigation was still applied under scarce photosynthesis. Substrate temperature measurement could be a new method for scanning real time irrigation in volcanic rock substrates. The fuzzy controller becomes an excellent option for plant growth management and healthy tomato production but shade control should be included.

REFERENCES

[1] J. Tanny, S. Cohen, and A. Grava, "Airflow and turbulence in a banana screenhouse," Acta Horticulturae, vol. 719, pp. 623-630, 2006.

[2] J. C. Bakker, "Russeting (cuticle cracking) in glasshouse tomatoes in relation to fruit growth," J. Hort. Sci, vol. 63, pp. 459–463, 1988.

[3] T. Wada, H. Ikeda, K. Matsushita, A. Kambara, H. Hirai, and K. Abe, "Effects of shading in summer on yield and quality of tomatoes grown on a single-truss system," Journal Japan Society Horticultural Science, vol. 75, pp. 51-58, 2006.

[4] H.F. Plaisier, and L. Svensson, "Use of adapted energy screens in tomato production with higher water vapour transmission," Acta Horticulturae, vol. 691, pp. 583-588, 2005.

[5] L.D. Albright, A.J. Both, and A. Chiu, "Controlling greenhouse light to a consistent daily integral," Transanctions of the ASAE, vol. 43, pp. 421-431, 2000.

[6] K. Kell, M. Beck, and F.W. Frenz, "Automated ecological fertilization and irrigation of soil grown crops in greenhouses with a computer controlled system (KLIWADU)," Acta Hortic., vol. 481, pp. 609-616, 1999.

[7] P. Javadi, A. Tabatabaee, M. Omid, R. Alimardani, and L. Naderloo, "Intelligent Control Based Fuzzy Logic for Automation of Greenhouse Irrigation System and Evaluation in Relation to Conventional Systems", Journal World Applied Sciences, vol. 6, pp. 16-23, 2009.

[8] F. Lafont, and J.-F Balmat, "Optimized fuzzy control of a greenhouse," Fuzzy Sets and Systems vol. 128, pp. 47–59, 2002.

[9] D. Gómez, A. López, G. Herrera, C. Fuentes, E. Rico, C. Olvera, D. Alaniz, T. Mercado, and S. Verlinde, "Fuzzy irrigation greenhouse control system based on a field programmable gate array," Afr. J. Agric. Res., Vol. 6, pp. 2544-2557, 2011.

[10] S. L.Warren, and T.E. Bilderback, "Timing of low pressure irrigation affects plant growth and water utilization efficiency," Journal Environmental Horticulture, vol. 20, pp. 184–188, 2002.

[11] F. Giuffrida, "Temperature of substrates in relation to through characteristics," Acta Horticulturae, vol. 559, pp. 647-654, 2001.

[12] P. Lorenzo, M.L. García, M.C. Sánchez-Guerrero, E. Medrano, I. Caparrós, and M. Giménez, "Influence of Mobile Shading on Yield, Crop Transpiration and Water Use Efficiency," Acta Horticulturae, vol. 719, pp. 471-478, 2006.

[13] F. Jobin-Lawler, K. Simard, A. Gosselin, and A.P. Papadopoulos, "The influence of solar radiation and Boron-Calcium fruit application on cuticle cracking of a winter tomato crop grown under supplemental lighting," Acta Horticulturae, vol. 580, pp. 235-239, 2002.

[14] M. Dorais, A.P. Papadopoulos, and A. Gosselin, "Influence of electrical conductivity management on greenhouse tomato fruit yield and fruit quality," Agronomie, vol. 21, pp. 367-383, 2001.

[15] F. Hahn, "Fuzzy controller decreases tomato cracking in greenhouses," Comput. Electron. Agric., vol. 77, pp. 21-27, 2011.

AUTHORS PROFILE

Dr. Federico Hahn

Professor and researcher of biosystems and irrigation technology at the Universidad de Chapingo, Mexico.

Dr. Hahn develops control equipment for farmers and producers. His embedded systems are working around Mexico and other countries. His main areas of expertise are sensors, instrumentation and automation.

Plethora of Cyber Forensics

N.Sridhar[1]
Research Scholar, Dept.of CS&SE
Andhra University, Visakhapatnam,
Andhra Pradesh, India,
neralla_sridhar@yahoo.com

Dr.D.Lalitha Bhaskari[2]
Associate Professor, Dept.of CS&SE
Andhra University, Visakhapatnam,
Andhra Pradesh, India,
lalithabhaskari@yahoo.co.in

Dr.P.S.Avadhani[3]
Professor, Dept.of CS&SE
Andhra University, Visakhapatnam,
Andhra Pradesh, India,
psavadhani@yahoo.com

***Abstract*— As threats against digital assets have risen and there is necessitate exposing and eliminating hidden risks and threats. The ability of exposing is called "cyber forensics." Cyber Penetrators have adopted more sophistical tools and tactics that endanger the operations of the global phenomena. These attackers are also using anti-forensic techniques to hide evidence of a cyber crime. Cyber forensics tools must increase its toughness and counteract these advanced persistent threats. This paper focuses on briefing of Cyber forensics, various phases of cyber forensics, handy tools and new research trends and issues in this fascinated area.**

Keywords- Cyber Forensics; digital evidence; forensics tools; cyber crimes.

I. INTRODUCTION

As Internet technologies proliferate into everyday life, we come close to realizing new and existing online opportunities. One such opportunity is in Cyber forensics, unique process of identifying, preserving, analyzing and presenting digital evidence in a manner that is legally accepted. The American Heritage Dictionary defines forensics as "relating to the use of science or technology in the investigation and establishment of facts or evidence in a court of law" [1].

Computer forensics involves the identification, documentation, and interpretation of computer media for using them as evidence and/or to rebuild the crime scenario [2]. According to [3] computer forensics defined as the process of identifying, collecting, preserving, analyzing and presenting the computer-related evidence in a manner that is legally acceptable by court.

More recently, computer forensics branched into several overlapping areas, generating various terms [4] such as, digital forensics, data forensics, system forensics, network forensics, email forensics, cyber forensics, forensics analysis, enterprise forensics, proactive forensics etc., as shown in figure-1. Digital forensics is the investigation of what happened and how. System forensics is performed on standalone machines. Network forensics involves the collection and analysis of network events in order to discover the sources of security attacks. The same process applied on Web is also known as Web forensics. Data forensics major focuses on analysis of volatile and non-volatile data. Proactive forensics is an ongoing forensics and there is an opportunity to actively, and regularly collect potential evidence in an ongoing basis. Email forensics deals with one or more e-mails as evidence in forensic investigation. Enterprise forensics is named in the context of enterprise; it is primarily concerned with incident response and recovery with little concern about evidence. Cyber forensics focuses on real-time, on-line evidence gathering.

Forensics analysis deals with identification, extraction and reporting on data obtained from a computer system.

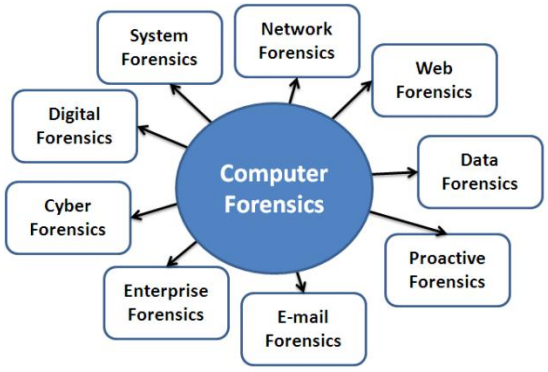


Figure 1: Various types of Computer Forensics

A. Cyber Crimes

In the 2008 CSI Computer Crime and Security Survey, it was noted that there is an average loss of $500,000 with corporations experiencing financial fraud (related to computing) and an extra average of $350,000 losses at companies that experienced "bot" attacks [5]. As per [6], the collecting digital devices in a forensically sound manner is becoming more critical since 80% of all cases have some sort of digital evidence involved in them.

With increased use of Internet in homes and offices, there has been a proliferation of cyber-related crimes and these crimes investigation is a tedious task. Cybercrime is typically described as any criminal act dealing with computers or computer Networks [7]. Cybercrimes can be classified into three groups [2]; Crimes directed against computer, crimes where the computer contains evidence, and crimes where the computer is used to commit the crime. Other names of cybercrime are e-crime, computer crime or Internet crime. Cybercrimes spread across the world with various names like worms, logic bombs, botnets, data diddling, mail bombing, phishing, rootkits, salami theft, spoofing, zombie, time bomb, Trojan horse etc.

Using the Internet, a person sitting in a Net cafe of a remote location can attack a computer resource in USA using

a computer situated in Britain as a launch pad for his attack. Challenges behind these situations are both technological and jurisdictional. Confidentiality, integrity and availability are the cardinal pillars of cyber security and they should not be compromised in any manner [2]. Attackers also begin using anti-forensic techniques to hide evidence of a cybercrime. They may hide folders, rename files, delete logs, or change, edit or modify file data [7]. To combat these kinds of crimes, Indian Government established Cyber Forensics Laboratory in November, 2003.

B. Overview of Cyber Forensics

Cyber forensics becoming as a source of investigation because human expert witnesses are important since courts will not recognize software tools such as Encase, Pasco, Ethereal as an expert witness [8]. Cyber forensics is useful for many professionals like military, private sector and industry, academia, and law. These areas have many needs including data protection, data acquisition, imaging, extraction, interrogation, normalization, analysis, and reporting.

It is important for all professionals working in the emerging field of cyber forensics to have a working and functioning lexicon of terms like bookmarks, cookies, webhit etc., that are uniformly applied throughout the profession and industry. Cyber forensics international guidelines, related key terms and tools are focused in the cyber forensics field manual [7].

The objective of Cyber forensics is to identify digital evidence for an investigation with the scientific method to draw conclusions. Examples of investigations that use cyber forensics include unlawful use of computers, child pornography, and cyber terrorism.

The area of cyber forensics has become prominent field of research because:

1) Forensics systems allow the administrator to diagnose errors
2) Intrusion detection systems are necessary in avoiding cyber crimes
3) Change detection can be possible with proactive forensics

Cyber forensics can be used for two benefits [9]:

1) To investigate allegations of digital malfeasance
2) To perform cause analysis

II. Phases of Cyber Forensics

Cyber forensics has four distinct phases: incident identification, acquisition of evidence, analysis of evidence, and reporting with storage of evidence [10]. Figure 2 shows various phases of cyber forensics process and each phase responsibility. The identification phase mainly deals with incident identification, evidence collection and checking of the evidence. The acquisition phase saves the state of a computer system that can be further analyzed. The analysis phase collects the acquired data and examines it to find the pieces of evidences. The reporting phase comprises of documentation and evidence retention.

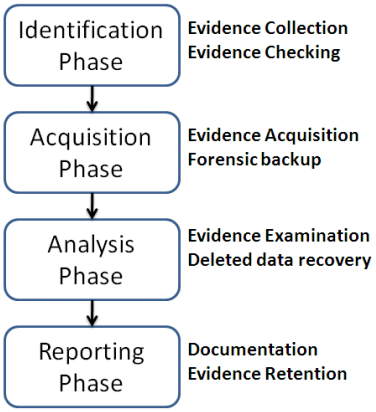


Figure 2: Phases of Cyber Forensics

A. Identification Phase

The identification phase is the process of identifying evidence material and its probable location. This phase is unlike a traditional crime scene it processes the incident scene and documents every step of the way. Evidence should be handled properly. Basic requirement in evidence collection is evidence must be presented without alteration. This requirement applies to all phases of forensics analysis. At the time of evidence collection, there is a need of thorough check of system logs, time stamps and security monitors.

Once evidence collected, it is necessary to account for its whereabouts. Investigators would need detailed forensics to establish a chain of custody, the documentation of the possession of evidence. Chain of custody is a vital part of computer forensics and the legal system [11] and the goal is to protect the integrity of evidence, so evidence should be physically secured in a safe place along with a detailed log. Figure 3 shows the evidence and chain of custody which is useful during incident investigation. Handling specific type of incidents like Denial of Service, Malicious Code, Unauthorized access etc are described in computer security incident handling guide [12].

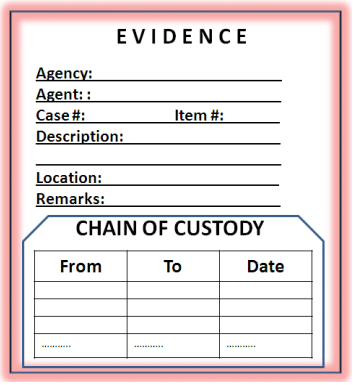


Figure 3: Evidence Form and Chain of Custody

B. Acquisition Phase

The acquisition phase saves the state of evidence that can be further analyzed. The goal of this phase is to save all digital values. Here, a copy of hard disk is created, which is commonly called as an image. Different methods of acquiring

data and their relative advantages and disadvantages are described in [13]. As per law enforcement community, there are three types of commonly accepted forensics acquisition: mirror image, forensics duplication and live acquisition.

Mirror images, bit-for-bit copy, involve the backups of entire hard disk. Creation of mirror image is simple in theory, but its accuracy must meet evidence standards. The purpose of having mirror image is evidence available in the case of the original system need to be restarted for further analysis.

A forensic duplicate, sector-by-sector, is an advanced method that makes a copy of every bit without leaving any single bit of evidence. The resultant may be single large file and must be an exact representation of the original drive at bit-stream level. This method is most common type of acquisition because it creates a forensic image of the e-evidence and it also contains file slack. In case of the small file overwrites a larger file, surplus bytes are available in the file slack. The forensic duplication process can be done with the help of tools like Forensic Tool Kit (FTK) imager, UNIX dd command, or Encase. Access Data's FTK is one of the powerful tools available and one of the promising features is the ability to identify steganography and practice of camouflaging data in plain sight.

It is often desirable to capture volatile information, which is stored in RAM; it cannot be collected after the system has been powered down. This information may not be recorded in a file system or image backup, and it may hold clues related to attacker. All currently running processes, open sockets, currently logged users, recent connections etc, are available in volatile information.

Generally, intruder takes steps to avoid detection. Trojans, keyloggers, worms etc., are installed in subtle places. One of such things to be considered in the acquisition process is rootkits, automated packages that create backdoors. An Intruders/hackers use rootkits to remove log files and other information to hide the presence of intruder. Mobile phones are become tools for cybercrimes, mobile phone evidence acquisition testing process are discussed in [14].

C. Analysis Phase

Forensic analysis is the process of understanding, re-creating, and analyzing arbitrary events that have gathered from digital sources [15]. The analysis phase collects the acquired data and examines it to find the pieces of evidences. This phase also identify that the system was tampered or not to avoid identification. Analysis phase examines all the evidence collected during collection and acquisition phases. There are three types of examinations can be applied for the forensics analysis; limited, partial or full examination.

Limited examination covers the data areas that are specified by legal documents or based on interviews. This examination process is the least time consuming and most common type. Partial examination deals with prominent areas. Key areas like log files, registry, cookies, e-mail folders and user directories etc., are examined in this case of partial examination. This partial examination is based on general search criteria which are developed by forensic experts. Most time consuming and less frequent examination process are full examination. This requires the examiner to look each, and every possible bit of data to find the root causes of the incident. File slack inspection is done in this examination.

Some of tools used in the analysis phase are Coroner, Encase, FTK. The Coroner toolkit run under UNIX and EnCase is a toolkit that runs under Windows [7]. EnCase has the ability to process larger amounts and allow the user to use predefined scripts to pull information from the data being processed. FTK contains a variety of separate tools (text indexing, NAT recovery, data extraction, file filtering and email recovery etc.) to assist in the examination.

D. Reporting Phase

The reporting phase comprises of documentation and evidence retention. The scientific method used in this phase is to draw conclusions based on the gathered evidence. This phase is mainly based on the Cyber laws and presents the conclusions for corresponding evidence from the investigation. There is a need of good policy for how long evidence from an incident should be retention. Factors to be considered in this process are prosecution, data retention and cost [12]. To meet the retention requirements there is a need of maintaining log archival [16]. The archived logs must be protected to maintain confidentiality and integrity of logs.

E. Forensics Methodology

The International Association of Computer Investigative Specialists (IACIS) has developed a forensic methodology which can be summarized as follows:

- ✓ Protect the Crime Scene, power shutdown for the computer and document the hardware configuration and transport the computer system to a secure location
- ✓ Bit Stream backup of digital media, use hash algorithms to authenticate data on all storage devices and document the system date and time
- ✓ Search keywords and check file space management (swap file, file slack evaluation, unallocated space)
- ✓ Evaluate program functionality, document findings/results and retain Copies of software

III. Cyber Forensics Tools

The main objective of cyber forensics tools is to extract digital evidence which can be admissible in court of law. Electronic evidence (e-evidence, for short) is playing a vital role in cybercrimes. Computer forensics tools used to find skeletons in digital media. To reduce the effect of anti-forensics tools the Investigator is likely to have the tools and knowledge required to counter the use of anti-forensics techniques [17]. Sometimes collection of digital evidence is straightforward because intruders post information about themselves from Facebook, Orkut, Twitter, MySpace and chat about their illegal activities. A subpoena, rather than special forensics tools, required obtain this information; these e-mails or chats from social networks can be admissible as evidence. A snapshot of the state of the art of forensic software tools for mobiles given in [18]. The process model for cellular phone tool testing had shown in [14]. Various cyber forensics tools and their description are provided in [7] some of them are:

1. The Coroner's Toolkit (TCT), is an open source set of forensic tools designed to conduct investigation UNIX systems.
2. Encase is the industry standard software used by law enforcement
3. The Forensic Toolkit (FTK) is very powerful tool but not simple to use.
4. I2Analyst is a different type of analysis tool, It is a visual investigative analysis software.
5. LogLogic's LX 2000 is powerful and distributed log analysis tool.
6. NetWitness and security intelligence, are network traffic security analyzer tools.
7. ProDiscover Incident Response (IR) is a complete IT forensic tool that can access computers over the network to study the network behavior
8. The Sleuth Kit is one of network forensics tools used to find file instances in an NTFS file

IV. Current Research

Cyber Forensics is sizzling topic of the current trends. Many researchers started doing intensive research in this current area. New directions in this field include authorship analysis, digital evidence collection and forensics investigation process, proactive forensics, intrusion detection systems with the help of honeypots, building evidence graphs, identifying usage of mobile phones in cybercrimes and hash function for preserving the integrity of evidence. The complete picture of Cyber Forensics in the form of Cyber forensics ontology which can be helpful for studying cyber forensics is given in [19]. Proactive forensics helps in the creation of preventive intelligence and threat monitoring besides post incident investigations.

Advantages and disadvantages of intercepting wireless network traffic as a means of locating potential evidence sources during evidence seizure are listed in [20]. Also in the same work the advantages and disadvantages of impairing communications to or from 802.11-based wireless networks during forensic seizure were discussed. High speed bitwise search model for large-scale digital forensic investigations using pattern matching board to search for string and complex regular expressions discussed in [21].

Various methods on how the evidential value of digital timestamps can be enhanced by taking a hypothesis based approach to the investigation of digital timestamp in his thesis work are proposed in the thesis [17]. Analysis of Instant Messaging in terms of computer forensics and intrusion detection is unexplored until now. Authorship classification used for forensics analysis or masquerade detection [22]. Creation of mobile software that runs on a mobile device and the goal is to aid crime scene personnel in the collection of digital devices during the course of an investigation is proposed by [23].

V. Conclusion

Cyber forensics is an emerging field in the 21st century. Detailed study of the field of cyber forensics is given in this paper. When analyzing cyber forensics, the process of doing so is different from the traditional forensics. In this paper, we described various computer forensics related definitions and phases cyber forensics and forensics methodology.

Various phases of the Cyber forensics have been discussed and each phase explored with their respective tools. Moreover, we mentioned different tools that are utilized in cyber forensics. Finally, we had shown the current research trends in this new era of cyber forensics; it still evolves and will remain a hot topic as long as there are ways to threaten data security.

References

[1] Kruse W.G, and Heiser J.G, *Computer Forensics Incident Response Essentials*, 2002, Addison Wesley Pearson Education, Boston

[2] Ibrahim M. Baggili, Richard Mislan, Marcus Rogers, *Mobile Phone Forensics Tool Testing: A Database Driven Approach*, International Journal of Digital Evidence Fall 2007, Volume 6, Issue 2

[3] Caloyannides, Michael A, Computer Forensics and Privacy. Artech House, Inc. 2001.

[4] Deepak Singh Tomar, Nikhil Kumar Singh, Bhopal Nath Roy, *An approach to understand the end user behavior through log analysis*, International Journal of Computer Applications (0975 – 8887) Volume 5– No.11, August 2010

[5] Svein Yngvar Willassen, Methods for Enhancement of Timestamp Evidence in Digital Investigations, Doctoral thesis at NTNU, 2008: 19

[6] Wayne Jansen, Rick Ayers, *Forensic Software Tools for Cell Phone Subscriber Identity Modules*, Conference on Digital Forensics, Security and Law, 2006

[7] Ashley Brinson, Abigail Robinson, Marcus Rogers, *A cyber forensics ontology: Creating a new approach to studying cyber forensics*, Digital Instigation, Elsevier, 2006

[8] Benjamin Turnbull, Jill Slay, *Wireless Forensic Analysis Tools for use in the Electronic Evidence Collection*, IEEE Proceedings of the 40th Annual Hawaii International Conference on System Sciences-2007 (HICSS'07)

[9] Hyungkeun Jee, Jooyoung Lee, and Dowon Hong, High Speed Bitwise Search for Digital Forensic System, Proceedings of world academy of science engineering and technology, volume 26, 2007.

[10] Angela Orebaugh and Jeremy Allnutt, *Classification of Instant Messaging Communications for Forensics Analysis*, The International Journal of Forensics Computer Science, 2009 (1), 22-28

[11] Ibrahim Baggili, Generating System Requirements for a Mobile Digital Device Collection System, European and Mediterranean Conference on Information Systems 2010, Abudhabi, UAE

Authors' Profile

N.Sridhar is a research scholar in Andhra University under the supervision of Prof.P.S.Avadhani and Dr.D.Lalitha Bhaskari. He received his M.Tech (IT) from Andhra University and presently working as Associate Professor in IT Department of GMRIT. He is a Life Member of ISTE. His research areas include Network Securiy, Cryptography, Cyber Forensics and Web Security.

Mrs. Dr. D. Lalitha Bhaskari is an Associate Professor in the department of Computer Science and Engineering of Andhra University. She is guiding more than 8 Ph. D Scholars from various institutes. Her areas of interest include Theory of computation, Data Security, Image Processing, Data communications, Pattern Recognition. Apart from her regular academic activities she holds prestigious responsibilities like Associate Member in the Institute of Engineers, Member in IEEE, Associate Member in the Pentagram Research Foundation, Hyderabad, India.

Dr. P. S. Avadhani is a Professor in the department of computer Science and Engineering of Andhra University. He has guided 7 Ph. D students, 3 students already submitted and right now he is guiding 12 Ph. D Scholars from various institutes. He has guided more than 100 M.Tech. Projects. He received many honors and he has been the member for many expert committees, member of Board of Studies for various universities, Resource person for various organizations. He has co-authored 4 books. He is a Life Member in CSI, AMTI, ISIAM, ISTE, YHAI and in the International Society on Education Technology. He is also a Member of IEEE, and a Member in AICTE.

A Fuzzy Similarity Based Concept Mining Model for Text Classification

Text Document Categorization Based on Fuzzy Similarity Analyzer and Support Vector Machine Classifier

Shalini Puri
M. Tech. Student
BIT, Mesra
India

***Abstract*—Text Classification is a challenging and a red hot field in the current scenario and has great importance in text categorization applications. A lot of research work has been done in this field but there is a need to categorize a collection of text documents into mutually exclusive categories by extracting the concepts or features using supervised learning paradigm and different classification algorithms. In this paper, a new Fuzzy Similarity Based Concept Mining Model (FSCMM) is proposed to classify a set of text documents into pre - defined Category Groups (CG) by providing them training and preparing on the sentence, document and integrated corpora levels along with feature reduction, ambiguity removal on each level to achieve high system performance. Fuzzy Feature Category Similarity Analyzer (FFCSA) is used to analyze each extracted feature of Integrated Corpora Feature Vector (ICFV) with the corresponding categories or classes. This model uses Support Vector Machine Classifier (SVMC) to classify correctly the training data patterns into two groups; i. e., + 1 and – 1, thereby producing accurate and correct results. The proposed model works efficiently and effectively with great performance and high - accuracy results.**

Keywords-Text Classification; Natural Language Processing; Feature Extraction; Concept Mining; Fuzzy Similarity Analyzer; Dimensionality Reduction; Sentence Level; Document Level; Integrated Corpora Level Processing.

I. INTRODUCTION

From the long time, the discipline of Artificial Intelligence (AI) is growing up on the map of science with psychology and computer science. It is an area of study that embeds the computational techniques and methodologies of intelligence, learning and knowledge [1] to perform complex tasks with great performance and high accuracy. This field is fascinating because of its complementarities of art and science. It contributes to increase the understanding of reasoning, learning and perception. Natural Language Processing (NLP) is the heart of AI and has text classification as an important problem area to process different textual data and documents by finding out their grammatical syntax and semantics and representing them in the fully structured form [1] [2]. AI provides many learning methods and paradigms to represent, interpret and acquire domain knowledge to further help other documents in learning.

Text Mining (TM) is a new, challenging and multi-disciplinary area, which includes spheres of knowledge like Computing, Statistics, Predictive, Linguistics and Cognitive Science [3] [4]. TM has been applied in a variety of concerns and applications. Some applications are summary creation, clustering, language identification [5], term extraction [6] and categorization [5] [6], electronic mail management, document management, and market research with an investigation [3] [4] [5].

TM consists of extracting regularities, patterns, and categorizing text in large volume of texts written in a natural language; therefore, NLP is used to process such text by segmenting it into its specific and constituent parts for further processing [2]. Text segmentation is also an important concern of TM. Many researches go on for the work of text classification [3] [4] [6] [7] [8] just for English language text. Text classification is performed on the textual document sets written in English language, one of the European Language, where words can be simply separated out using many delimiters like comma, space, full stop, etc. Most of the developed techniques work efficiently with European languages where the words, the unit of representation, can be clearly determined by simple tokenization techniques. Such text is referred as segmented text [10]. It does not always happen. There are some Asian Languages in which textual document does not follow word separation schemes and techniques. These languages contain un-segmented text. There are so many other languages in the world like Chinese and Thai Languages in which there is no delimiter to separate out the words. These languages are written as a sequence of characters without explicit word boundary delimiters [10]. So, they use different mechanisms for segmentation and categorization. Here, the proposed effort is to work only on the English text documents.

Therefore, TM categorization is used to analyze and comprise of large volume of non - structured textual documents. Its purpose is to identify the main concepts in a text document and classifying it to one or more pre-defined categories [3]-[12]. NLP plays an important and vital role to convert unstructured text [4] [5] [6] into the structured one by performing a number of text pre-processing steps. This processing results into the extraction of specific and exclusive

concepts or features as words. These features help in categorizing text documents into classified groups.

Section II discusses the thematic background and related research work done on the concept mining, feature extraction, and similarity measure. In section III, the proposed model and methodology is discussed in detail. It discusses Text Learning Phase which includes Text Document Training Processor (TDTP), Pseudo Thesaurus, Class Feeder (CF) and Fuzzy Feature Category Similarity Analyzer (FFCSA). Next, Support Vector Machine Classifier (SVMC) and Text Classification Phase are discussed. Finally, section IV concludes the paper with the suggested future work and scope of the paper.

II. Thematic Background And Related Work

Supervised learning techniques and methodologies for automated text document categorization into known and predefined classes have received much attention in recent years. There are some reasons behind it. Firstly, in the unsupervised learning methods, the document collections have neither predefined classes nor labeled document's availability. Furthermore, there is a big motivation to uncover hidden category structure in large corpora. Therefore, text classification algorithms are booming up for word based representation of text documents and for text categorization.

A. Concept Mining

Concept Mining is used to search or extract the concepts embedded in the text document. These concepts can be either words or phrases and are totally dependent on the semantic structure of the sentence. When a new text document is introduced to the system, the concept mining can detect a concept match from this document to all the previously processed documents in the data set by scanning the new document and extracting the matching concepts [5]. In this way, the similarity measure is used for concept analysis on the sentence, document, and corpus levels.

These concepts are originally extracted by the semantic role labeler [5] and analyzed with respect to the sentence, document, and corpus levels. Thus, the matching among these concepts is less likely to be found in non - related documents. If these concepts show matching in unrelated documents, then they produce errors in terms of noise. Therefore, when text document similarity is calculated, the concepts become insensitive to noise.

B. Feature Extraction

In text classification, the dimensionality of the feature vector is usually huge. Even more, there is the problem of *Curse of Dimensionality*, in which the large collection of features takes very much dimension in terms of execution time and storage requirements. This is considered as one of the problems of *Vector Space Model (VSM)* where all the features are represented as a vector of *n* - dimensional data. Here, *n* represents the total number of features of the document. This features set is huge and high dimensional.

There are two popular methods for feature reduction: *Feature Selection and Feature Extraction*. In feature selection methods, a subset of the original feature set is obtained to make the new feature set, which is further used for the text classification tasks with the use of Information Gain [5]. In feature extraction methods, the original feature set is converted into a different reduced feature set by a projecting process. So, the number of features is reduced and overall system performance is improved [6].

Feature extraction approaches are more effective than feature selection techniques but are more computationally expensive. Therefore, development of scalable and efficient feature extraction algorithms is highly demanded to deal with high-dimensional document feature sets. Both feature reduction approaches are applied before document classification tasks are performed.

C. Similarity Measure

In recent years, fuzzy logic [1] [2] [4] [6]-[15] has become an upcoming and demanding field of text classification. It has its strong base of calculating membership degree, fuzzy relations, fuzzy association, fuzzy production rules, fuzzy k-means, fuzzy c-means and many more concerns. As such, a great research work has been done on the fuzzy similarity and its classifiers for text categorization.

The categorizer based on fuzzy similarity methodology is used to create categories with a basis on the similarity of textual terms [4]. It improves the issues of linguistic ambiguities present in the classification of texts. So, it creates the categories through an analysis of the degree of similarity of the text documents that are to be classified. The similarity measure is used to match these documents with pre-defined categories [4] - [15]. Therefore, the document feature matrix is formed to check that a document satisfies how many defined features of the reduced feature set and categorized into which category or class [4] [6] [7]. The fuzzy similarity measure can be used to compute such different matrices.

III. The Fuzzy Similarity Based Concept Mining Model (FSCMM)

In this section, the proposed *Fuzzy Similarity Based Concept Mining Model (FSCMM)* is discussed. This model automatically classifies a set of known text documents into a set of category groups. The model shows that how these documents are trained step by step and classified by the Support Vector Machine Classifier (SVMC). SVMC is further used to classify various new and unknown text documents categorically.

The proposed model is divided into the two phases*: Text Learning Phase (TLP)* and *Text Classification Phase (TCP)*. TLP performs the learning function on a given set of text documents. It performs the steps of first stage; i. e., *Text Document Training Processor (TDTP)* and then the steps of second stage; i. e., *Fuzzy Feature Category Similarity Analyzer (FFCSA)*. The TDTP is used to process the text document by converting it into its small and constituent parts or chunks by using NLP concepts at the *Sentence, Document and Integrated Corpora Level*s. Then, it searches and stores the desired, important and non-redundant concepts by removing stop words, invalid words and extra words. In the next step, it performs word stemming and feature reduction. The result of sentence level preparation is low dimensional *Reduced Feature Vector (RFV)*. Each RFV of a document is sent for document

level preparation, so that *Integrated Reduced Feature Vector (IRFV)* is obtained. To obtain IRFV, all the RFVs are integrated into one. Now, *Reduced Feature Frequency Calculator (RFFC)* is used to calculate the total frequency of each different word occurred in the document. Finally, all redundant entries of each exclusive word are removed and all the words with their associated frequencies are stored in decreasing order of their frequencies. At the integrated corpora level, the low dimension *Integrated Corpora Feature Vector (ICFV)* is resulted.

In such a way, feature vectors at each level are made low dimensional by processing and updating step by step. Such functionality helps a lot to search the appropriate concepts with reduced vector length to improve system performance and accuracy.

FFCSA performs similarity measure based analysis for feature pattern (TD – ICFV) using the enriched fuzzy logic base. The membership degree of each feature is associated with it. Therefore, an analysis is performed between every feature of a text document and class.

SVMC is used for the categorization of the text documents. It uses the concept of hyper planes to identify the suitable category. Furthermore, SVMC accuracy is checked by providing some new and unknown text documents to be classified into the respective Category Group (CG). This task is performed in TCP.

The proposed Fuzzy Similarity Based Concept Mining Model (FSCMM) is shown in Fig. 1. In the next sections, this model and its components are discussed in detail.

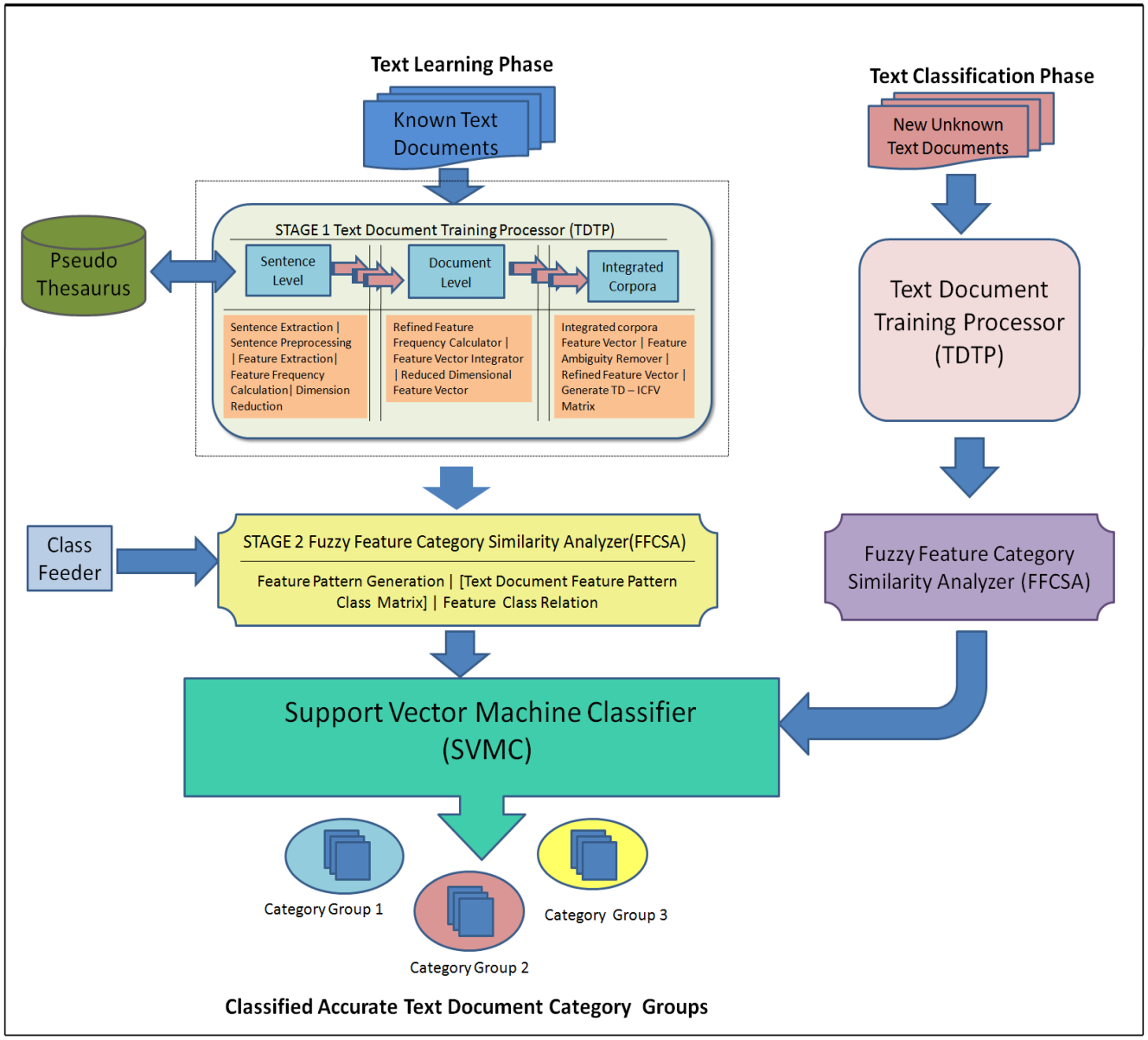


Figure 1. The Fuzzy Similarity Based Concept Mining Model (FSCMM).

A. Text Learning Phase (TLP)

Consider a set of *n* text documents,

$$TD = \{TD_1, TD_2, TD_3, \ldots, TD_n\} \quad (1)$$

Where TD_1, TD_2, $TD_3, \ldots, TD_n$ are the individual and independent text documents which are processed, so that they can be categorized into the required category.

1) Text Document Training Processor (TDTP)

Text Document Training Processor (TDTP) prepares the given text document set *TD* of *n* text documents by performing many operations on the sentence, document, and integrated corpora levels. Firstly, each text document TD_i, $1 \leq i \leq n$, is processed at its sentence level. The result of such sentence level pre-processing for all the sentences of TD_i is integrated into one, which is further processed and refined to make available for the integrated corpora. Integrated corpora accept and integrate all the refined text documents and perform more processing. Its result is sent to FFCSA.

a) At Sentence Level

This section describes that how a sentence is pre-processed and finds out the feature vector, each feature's frequency and finally, the conversion of the *Feature Vector (FV)* into the *Reduced Feature Vector (RFV).*

A text document TD_i is composed of a set of sentences, so consider

$$TD_i = \{s_{i1}, s_{i2}, s_{i3}, \ldots, s_{im}\} \quad (2)$$

Where *i* denotes the text document number and *m* denotes the total number of sentences in TD_i. *Sentence Extractor (SE)* is used to extract the sentence s_{i1} from *TDi*. Each sentence has its well-defined and non-overlapping boundaries, which makes the sentence extraction a simple task for SE.

When the sentence is extracted, a verb – argument structure is made for s_{i1}. The syntax tree is drawn using the pre-defined syntactic grammar to separate the verbs and the arguments of the sentence. The sentence can be composed of the nouns, proper nouns, prepositions, verbs, adverbs, adjectives, articles, numerals, punctuation and determiners. So, with the construction of the syntax tree, the stop word and other extra terms are removed except the nouns, proper nouns and numerals which are considered as the concepts. To remove the invalid and extra words, the *Pseudo Thesaurus* is used. It also helps in word stemming.

The next step is to make the *Feature Vector (FV)* of the sentence s_{ij} of text document TD_i as

$$FV = \{F_{i11}, F_{i12}, F_{i13}, \ldots, F_{i1r}\} \quad (3)$$

Where $1 \leq i \leq n$, $1 \leq j \leq m$, and *r* depicts the total number of present features in the s_{ij}. *Feature Frequency Calculator (FFC)* calculates the frequency of each different feature occurred in FV. Frequency represents the number of occurrences of a feature in the sentence. So, each different feature is associated with its frequency in the form of a *Feature Frequency pair as* $<F_{ijk}, freq\ (F_{ijk})>$, where *i* is the text document number, *j* is the sentence number of the sentence, *k* is the feature number, *freq ()* is a function to calculate the frequency of a feature, and $1 \leq k \leq r$.

The next step is to convert the high dimensional FV into low dimensional *Reduced Feature Vector (RFV*) to reduce the storage and execution time complexities. So, a counter loop is invoked to remove the duplicate or redundant entries of a feature. Therefore, only one instance of each different feature occurred is stored in RFV. It highly reduces the FV dimension and increases the efficiency of the system with good performance. The complete sentence level processing is shown in the Fig. 2.

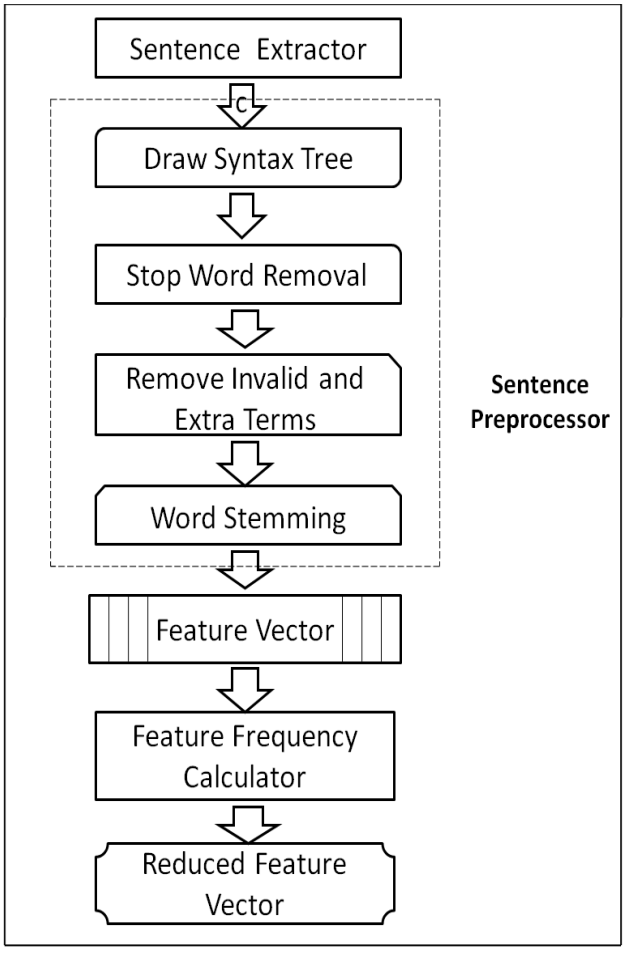


Figure 2. Text Document Preparation at Sentence Level.

Such sentence level preprocessing for the TDP is performed for each sentence of each document and then progressing toward the integrated corpora.

b) At Document Level

This step accepts the resultant RFV which is sent for document level pre-processing. Firstly, a counter loop is invoked to match the similar features in each of the RFV_j with every other RFV_q of a TD_i where $1 \leq j, q \leq m, j \neq q$ and $1 \leq i \leq n$. *Refined Feature Vector Calculator (RFVC)* updates each feature's frequency for those features which are present two or more times in more than two sentences. These updates are done by adding up their frequencies in terms of combined calculated frequency of that feature only. In this way, it updates the count of each different occurred feature with more than one occurrence. The features that have occurred only once in the document will not update their frequency.

The next step is that all RFVs of a TD_i are integrated into one as

$$IRFV = Integrat\ (RFV_1, RFV_2, \ldots, RFV_m) \quad (4)$$

Where *Integrat ()* is a function to combine all RFVs.

Now each RFV_j is compared with every other RFV_q where $1 \leq j, q \leq m$, and $j \neq q$. In this way, each feature of RFV_j is compared with the every other feature of the RFV_q and thereby, the duplicate and redundant features are removed. The complete procedure on the document level is shown in Fig. 3.

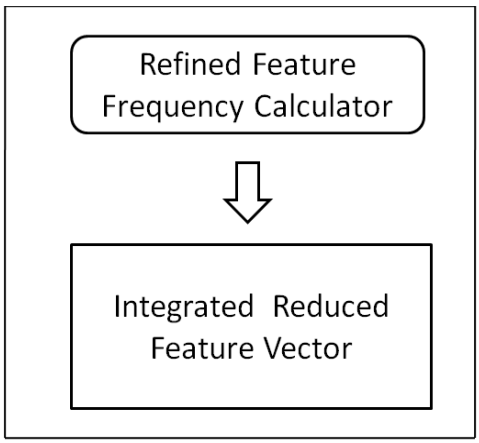


Figure 3. Text Document Preparation at Document Level.

c) At Integrated Corpora

In integrated corpora, all the IRFVs of *n* documents are integrated into one. This step is used to calculate and update the final frequency of each different feature occurred in the corpora. Firstly, it removes duplicate and redundant feature entries of IRFV, and then removes all ambiguous words with the help of the pseudo thesaurus. In such a way, *Integrated Corpora Feature Vector (ICFV)* is generated. A *Threshold Value (TV)* is defined for ICFV. TV cuts off those features whose *total frequency is less than TV*. Finally, it reduces the dimension of ICFV.

Therefore, *Integrated Corpora Feature Vector (ICFV)* is constructed as

$$ICFV = \{F_1, F_2, \ldots, F_n\} \quad (5)$$

Where $F_1, F_2, \ldots, F_n$ represent the features.

Such features with their associated frequencies show the statistical similarity between the text documents. Each feature is counted for each document and represented as the feature and its frequency as follows.

$$TD_1 = \{<F_1, f(F_1)>, <F_2, f(F_2)>, \ldots, <F_n, f(F_n)>\} \quad (6)$$

Where $f(F_i)$ is the function to calculate the frequency of feature F_i in the text document.

The next step is to generate the matrix of TD and ICFV in the given form of table 1. In this matrix, the 0 represents the absence of that feature in TD and the numerical value represents the total number of occurrences of the feature in the TD.

TABLE I. TD ICFV MATRIX

Text Document	Feature		
	F1	***F2***	***F3***
TD1	1	0	1
TD2	1	3	0
TD3	0	2	1
TD4	4	0	0

2) Pseudo Thesaurus

The Pseudo Thesaurus is a predefined English Vocabulary Set which is used to check the invalid words or to remove extra words from a sentence while processing the sentence in the TDTP. It is also used for word stemming so that the exact word can be obtained. For example, consider three different words for the word *research* - researching, researcher and researches. When the word stemming is performed, *research* is the final resulting word with the feature frequency counted as 3.

3) Class Feeder (CF)

Text Classification is the process of assigning the name of the class to a particular input, to which it belongs. The classes, from which the classification procedure can choose, can be described in many ways. So classification is considered as the essential part of many problems solving tasks or recognition tasks. Before classification can be done, the desired number of classes must be defined.

4) Fuzzy Fetaure Category Similarity Analyzer (FFCSA)

In FFCSA, firstly the *Feature Pattern (FP)* is made in the form of the membership degree of each feature with respect to every class. Consider text document set *TD* of *n* text documents as per given in equation 1, together with the ICFV *F* of *y* features $f_1, f_2, \ldots f_y$ and *e* classes $c_1, c_2, \ldots, c_e$. To construct the FP for each feature f_k in F, its FP fp_i is defined, by

$$fp_i = < fp_{i1}, fp_{i2}, fp_{i3}, \ldots, fp_{ie}> \quad (7)$$
$$= <\mu(f_i, c_1), \mu(f_i, c_2), \mu(f_i, c_3), \ldots, \mu(f_i, c_e)>$$

Where,

f_i represents the number of occurrences of f_i in the text document TD_g where $1 \leq g \leq n$.

$\mu(f_i, c_e)$ is defined as the sum of product of the feature value of f_i present in *n* text documents TD, w. r. t. a column vector and the 1 or 0 as the presence or absence of that feature in class c_e / Sum of the feature value for the class c_e only, where $1 \leq j \leq e$. So,

$$\mu(f_i, c_e) = \frac{\sum_g (TDgi).bi}{\sum_g (TDgi)} \quad (8)$$

b_i is represented as

$$b_i = 1, \text{ if } document\ \epsilon\ class\ c_e \quad (9)$$
$$= 0, otherwise$$

Each text document *TD* belongs to only one class *c*. In this way, each class can belong to one or more text documents. A

set of n documents and their related categories or classes are represented as an ordered pair as shown

$$TD = \{<TD_1, C(TD_1)>, <TD_2, C(TD_2)>, \ldots, <TD_m, C(TD_m)>\} \quad (10)$$

Where the class of the text document TD_i: $C(TD_i) \in C$, $C(TD)$ is a categorization function whose domain is TD and range is C.

Each document belongs to one of the classes in the $C(TD)$.The resulted text documents that have many features are stored with their relevant classes as shown in the table 2. They are in the form of *<doc no, number of occurrences of each feature, class no>*.

TABLE II. TEXT DOCUMENT FEATURE VECTOR CLASS MATRIX

Text Document	Feature			Class
	F1	*F2*	*F3*	
TD1	1	0	1	C1
TD2	1	3	0	C2
TD3	0	2	1	C2
TD4	4	0	0	C3

In such a way, the relation between a feature and a class is made. Sometimes, it is quite possible that one document belongs to two or more classes that concern has to be considered by making the more presences of the text document in the table with the cost of increased complexity, so it is required to check each feature's distribution among them.

B. Support Vector Machine Classifier (SVMC)

The next step is to use the *Support Vector Machine Classifier (SVMC*). SVMC is a popular and better method than other methods for text categorization. It is a kernel method which finds the maximum margin hyper plane in the feature space paradigm separating the data of training patterns into two groups like Boolean Logic 1 and 0. If any training pattern is not correctly classified by the hyper plane, then the concept of slack measure is used to get rid out of it.

Using this idea, SVMC can only separate apart two classes for $h = +1$ *and* $h = -1$. For e classes, one SVM for each class is created. For the SVM of class c_l, $1 \leq l \leq e$, the training patterns of class c_l are for the $h = +1$ and of other classes are $h = -1$. The SVMC is then the aggregation of these SVMs.

SVM provides good results then KNN method because it directly divide the training data according to the hyper planes.

C. Text Classification Phase

To check the predictive accuracy of the SVMC, new and unknown text document is used, which is independent of the training text documents and is not used to construct the SVMC. The accuracy of this document is compared with the learned SVMC's class. If the accuracy of the SVMC is acceptable and good, then it can be used further to classify the future unseen text documents for which the class label is not known. Therefore, they can be categorized into the appropriate and a suitable category group.

IV. CONCLUSION AND FUTURE SCOPE

Text classification is expected to play an important role in future search services or in the text categorization. It is an essential matter to focus on the main subjects and significant content. It is becoming important to have the computational methods that automatically classify available text documents to obtain the categorized information or groups with greater speed and fidelity for the content matter of the texts.

As the proposed FSCMM model is made for text document categorization, it works well with high efficiency and effectiveness. Although this model and methodology seem very complex, yet it achieves the task of text categorization with high performance, and good accuracy and prediction. Feature Reduction is performed on the sentence, document and integrated corpora levels to highly reduce feature vector dimension. Such reduction improves the system performance greatly in terms of space and time. Result shows that the feature reduction reduces the space complexity by20%.

Fuzzy similarity measure and methodology are used to make the matching connections among text documents, feature vectors and pre-defined classes. It provides the mathematical framework for finding out the membership degrees as feature frequency.

This model shows better results than other text categorization techniques. SVM classifier gives better results than KNN method. The system performance does not show high information gain and prediction results when KNN is used because it produces noise sensitive contents.

In the future, such model can be further extended to include the non-segmented text documents. It can also be extended to categorize the images, audio and video - related data.

REFERENCES

[1] N. P. Padhy, Artificial Intelligence and Intelligent Systems, 5th ed., Oxford University Press, 2009.

[2] Eliane Rich, Kevin Knight and Shivashankar B Nair, Artificial Intelligence, 3rd ed., Mc Graw Hill, 2010.

[3] Jiawei Han, and MicheLine Kamber, Data Mining: Concepts and Techniques, 2nd ed., Elsevier, 2006.

[4] Marcus Vinicius, C. Guelpeli, Ana Cristina, and Bicharra Garcia, "An Analysis of Constructed Categories for Textual Classification Using Fuzzy Similarity and Agglomerative Hierarchical Methods," Third International IEEE Conference Signal-Image Technologies and Internet-Based System, September 2008.

[5] S. Shehata, F. Karray, and M. S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 10, October 2010.

[6] Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee, "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification," IEEE Transactions On Knowledge And Data Engineering, Vol. 23, No. 3, March 2011.

[7] Choochart Haruechaiyasak, Mei-Ling Shyu, Shu-Ching Chen, and Xiuqi Li, "Web Document Classification Based on Fuzzy Association," IEEE, 2007.

[8] Ahmad T. Al-Taani, and Noor Aldeen K. Al-Awad, "A Comparative Study of Web-pages Classification Methods using Fuzzy Operators

Applied to Arabic Web-pages," World Academy of Science, Engineering and Technology, 2005.

[9] Qing YANG, Wei CHEN, and Bin WEN, "Fuzzy Ontology Generation Model using Fuzzy Clustering for Learning Evaluation," 2008.

[10] Kok WaiWong, Todsanai Chumwatana, and Domonkos Tikk, "Exploring The Use of Fuzzy Signature for Text Mining," IEEE, 2010.

[11] O. Dehzangi, M. J. Zolghadri, S. Taheri and S.M. Fakhrahmad, "Efficient Fuzzy Rule Generation: A New Approach Using Data Mining Principles and Rule Weighting," Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2007.

[12] Surya Sumpeno, Mochamad Hariadi, and Mauridhi Hery Purnomo, "Facial Emotional Expressions of Life-like Character Based on Text Classifier and Fuzzy Logic," IAENG International Journal of Computer Science, May, 2011.

[13] Rahmah Mokhtar, Siti Norul Huda Sheikh Abdullah, and Nor Azan Mat Zin, "Classifying Modality Learning Styles based on Production Fuzzy Rules," International Conference on Pattern Analysis and Intelligent Robotics, June 2011.

[14] Giuseppe Fenza, Vincenzo Loia, and Sabrina Senatore, "Concept Mining of Semantic Web Services By Means Of Extended Fuzzy Formal Concept Analysis (FFCA)," IEEE, Feb. 2008.

[15] Chen Yanyun , Qiu Jianlin, Gu Xiang,Chen Jianping, Ji Dan, and Chen Li, "Advances in Research of Fuzzy C-means Clustering Algorithm," International Conference on Network Computing and Information Security, 2011.

AUTHORS PROFILE

Shalini Puri is pursuing M.Tech in Computer Science at Birla Institute of Technology, Mesra, Ranchi, India. She is currently working as an Assistant Professor in a reputed engineering college in India. Her research areas include Artificial Intelligence, Data Mining, Soft Computing, Graph Theory, and Software Engineering.

Improved Echo cancellation in VOIP

Patrashiya Magdolina Halder
Dept. of Electronics and Telecommunication Engineering
Daffodil International University
Dhaka, Bangladesh

A.K.M. Fazlul Haque
Dept. of Electronics and Telecommunication Engineering
Daffodil International University
Dhaka, Bangladesh

***Abstract*— VoIP (voice over internet protocol) is very popular communication technology of this century and has played tremendous role in communication system. It is preferred by all because it deploys many benefits it uses Internet protocol (IP) networks to deliver multimedia information such as speech over a data network. VoIP system can be configured in these connection modes respectively; PC to PC, Telephony to Telephony and PC to Telephony. Echo is very annoying problem which occurs in VoIP and echo reduces the voice quality of VoIP. It is not possible to remove echo 100% from echoed signal because if echo is tried to be eliminated completely then the attempt may distort the main signal. That is why echo cannot be eliminated echo perfectly but the echo to a tolerable range. Clipping is not a good solution to suppress echo because part of speech may erroneously removed. Besides an NLP does not respond rapidly enough and also confuses the fading of the voice level at the end of a sentence with a residual echo. This paper has proposed echo cancellation in VoIP that has been tested and verified by MATLAB. The goal was to suppress echo without clipping and distorting the main signal. By the help of MATLAB program the echo is minimized to enduring level so that the received signal seems echo free. The percentage of suppressing echo varies with the amplitude of the main signal. With regarding the amplitude variation in received (echo free) signal the proposed method performs better in finding the echo free signal than the other conventional system.**

Keywords- PSTN; round trip delay; impedance; inverse filtering; denoise; histogram amplifier; repeater.

I. INTRODUCTION

Around 20 years of research on VoIP, some problems of VoIP are still remaining and a substantial problem in telecommunications is the generation of echo. Echo is a phenomenon where a delayed and distorted version of an original signal is reflected back to the source. Echo is a congenital problem which mainly occurs in PSTN (Public Switching Telephone Network) [1]. Echo occurs in analogy part of a telecommunication system .Echo is generated by human voice is heard as they are reflected from the floor, walls and other neighboring objects. If a reflected wave arrives after a very short time of direct sound, it is considered as a spectral distortion or reverberation. However, when the leading edge of the reflected wave appears again a few tens of milliseconds after the direct sound then it is heard as an audible echo [2]. Echo is annoying when the round trip delay exceeds 30 ms. such an echo is typically heard as a hollow sound. Echoes must be loud enough to be heard. Echo which is less than thirty 30 dB is rarely to be noticed. But when round trip delay exceeds 30 ms and echo strength exceeds 30 dB, echoes become steadily more disruptive. Every echo does not reduce voice quality. There are mainly two kinds of echo, that is Hybrid echo and Acoustic echo. Hybrid echo, line echo or electrical echo is different names given to the echo generated by an impedance mismatch in the analog local loop. The impedance mismatch occurs when the two-wire network meets the four-wire network [3]. The impedance of subscriber lines vary from one subscriber to the next, this time sharing makes it impossible to provide a perfect impedance match for every line [4]. Acoustic echo is caused by acoustic coupling problems between a telephone's speaker and its microphone. Acoustic echo can occur in mobile phones, wire line telephones or in a hands-free set of a speaker phone [4]. It can be caused by hand-set crosstalk in poor quality handsets or by echo in the environment surrounding the caller. There are some works on frequency reuse scheme [3, 6-9].

According to asterisk echo cancellation previously called carbon profile [6, 7] is operated by generating multiple copies of the received signal, each delayed by some small time increment. These delayed copies are then scaled and subtracted from the original received signal.Srinivasaprasath Raghavendran et al [3] has proposed an echo cancellation process using MATLAB but there the far end signal and the near end signal is taken separately and then tested whether there is echo or not by Double talk detector. This process also includes NLMS and subtraction .Ganesan Periakarruppan, Andy L.Y.Low, Hairul Azhar b Abdul Rashid et al [8] introduced that PBEC the sample to generate the echo replica model will be used to subtract the. Jerker Taudien el al [9] suggested Line probing is a method of inserting a known signal at the far-end and recording the near-end signal. The two signals are then analyzed together for various impediments. Three tone sweeps of different power levels are used to probe the line in the non-linear distortion analysis tool. The tone sweeps are recorded in three different power levels to detect clipping.

All these procedures require more things than this proposed method. This paper has suggested cancelling echo from echoed signal. Only received signal is analysed here so it is not needed to analyse near end and far end signal separately. Besides subtraction and clipping is not required here which ma affect the main signal. This is a very simple program which eliminates echo. Inverse filtering is used here which analyse the received signal and remove echo from the acquired signal.

II. BACKGROUND

Within the caller's telephone, a certain amount of the signal from the microphone is fed straight back to the earpiece. An improperly balanced hybrid won't correctly filter out the entire transmitted signal, and will reflect some of it back down the other half of the trunk. Imbalance may be from poor design (common) or unpredictable. The reasons of echo are as follows:

1. Poor room acoustics
2. Marginal microphones for soft terminals
3. Low quality cellular handsets
4. Deficient echo control in the terminal device its
5. Bridge-taps (something done by the Telco, seldom seen any more)
6. Use of lengthy untwisted wire within the subscriber's premises

III. SIMULATION AND RESULTS

This program is to explore the problem of echo cancellation. Iinverse filtering **(IF)** is a widely known method for voice and speech analysis, which mainly works on estimating the source of voiced speech. This method enables to estimate the glottal volume velocity waveform or glottal airflow. The idea behind inverse filtering is to form a computational model for glottal pulse detection by filtering the speech signal.

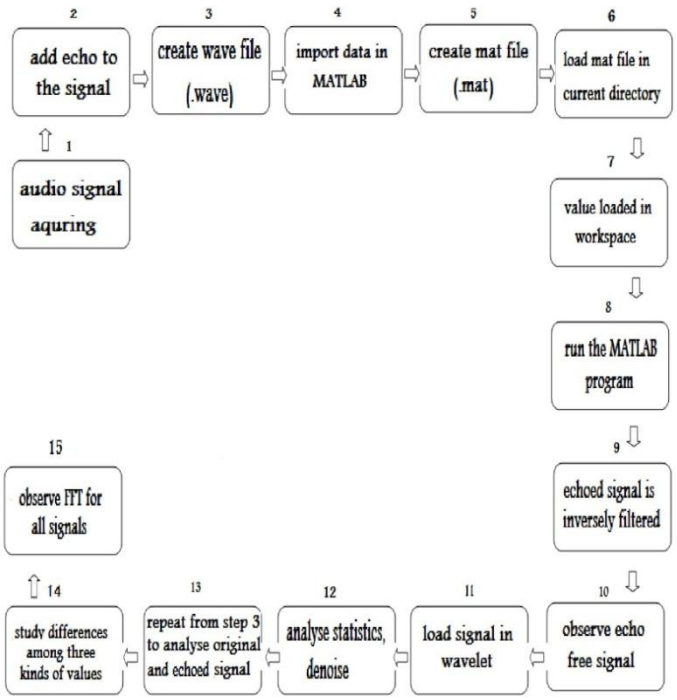


Figure 1. Flowchart for echo cancellation using MATLAB

This flowchart is given bellow for echo suppression using MATLAB (Fig. 1) .The main signal, signal with echo and signal without echo is showed respectively in Fig. 2, 3 and 4. This signal can be denoised using Matlab to remove echo from this signal.

For the experiment at first voice signal is acquired, it can be done by any kind of speech recorder. Then this acquired voice signal is used to create a **.wave** file using the audio signal. After that Matlab software is opened the data of the signal is imported to create **.mat** file. This mat file has to be loaded to transfer the value of the speech signal in workspace. To remove echo from the signal wavelet is used. To inspect the signals by wavelet wave menu is typed in the command window of Matlab; a new window of wavelet will be displayed then. The echoed signal is loaded in wavelet and then the signal is analyze, view the statistics, denoise and analyze the signal. Doing all these signal with echo can be analyzed. The signals can be compressed to get better result

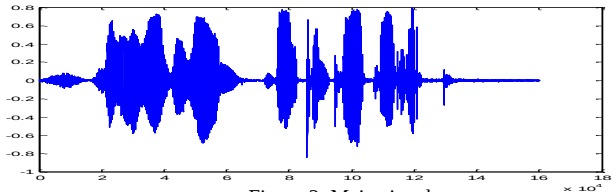


Figure 2. Main signal

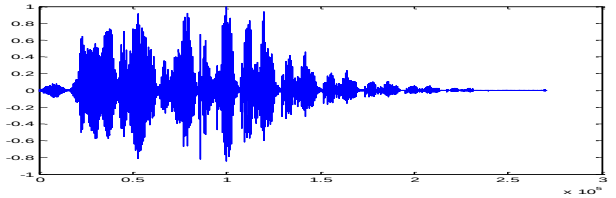


Figure 3. Signal with echo

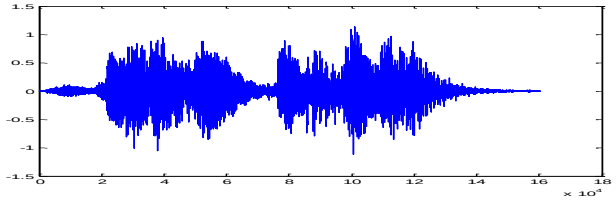


Figure 4. Signal without echo

By running the program, hearing the signal it will be clear that echo is removed from the echoed signal and the quality of signal is improved. The signal is clear and each part of speech is present. If the simulation figures are observed closely then it can be seen that the wave of main signal is like overlapped in the echoed signals figure. The refined signal is almost look alike the echo free signal, there is just change in the amplitude of the signal. So it can be said that the improved signal and the outcome signal both prove that the attempt is successful to remove echo and perform better.

Histogram is used to analyze the speech signal which showed in Fig. 5, 6 and 7.

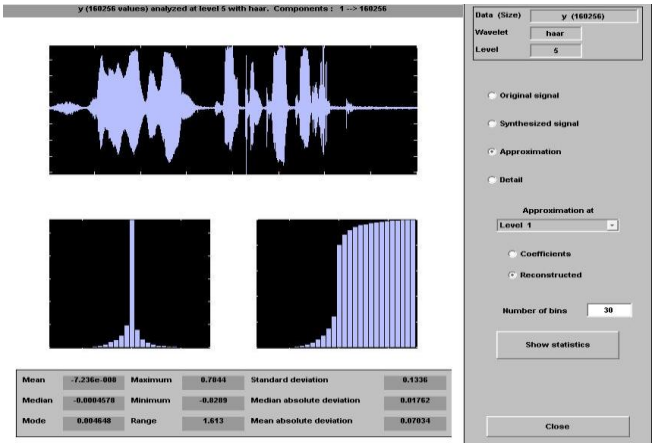


Figure 5. Analyzing main signal with wavelet

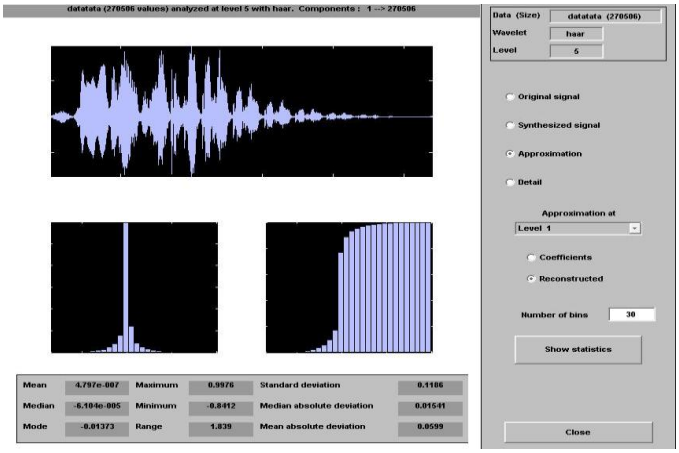


Figure 6. Analyzing signal with echo with wavelet

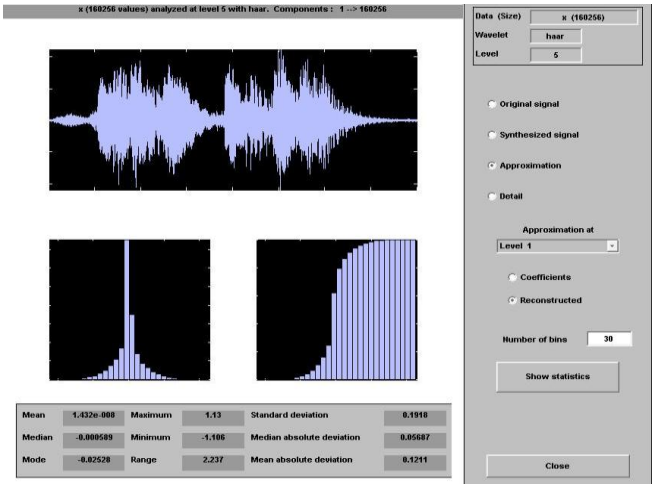


Figure 7. Analyzing signal without echo with wavelet

The following figure 8 shows the difference among the signals with the help of FFT (Fast Fourier Transform)

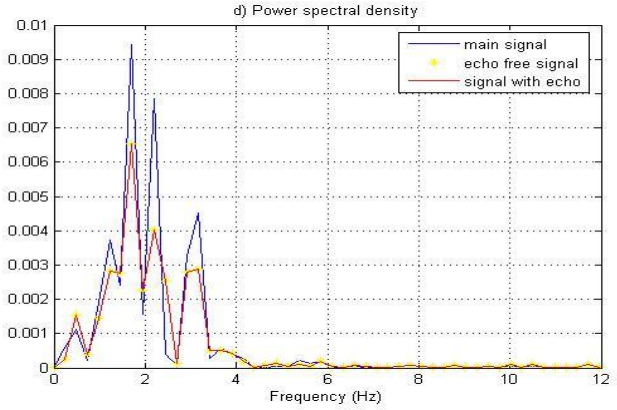


Figure 8. Analyzing all signals using FFT

Difference between signal with echo and signal without echo is observed. By seeing the waveforms generally it can be said that echo is removed. There is difference in amplitude among the signals; the amplitude can be regained by using amplifier or repeater.

For further analyzing different statistical values are extracted from this experiment and taking the values a table 4.4 and a graph (Fig. 9) is plotted, this will help to prove that echo is removes from the echoed signal.

TABLE I. SIGNAL ANALYZING

	Main signal		**Echoed signal**		**Signal without echo**	
Mean	-7.236e-008	0.7844	4.797e-007	0.999	1.432e-008	1.13
Median	-0.0004578	-0.8289	-6.104e-005	-0.8412	-0.000589	-1.106
Mode	0.000448	1.613	-0.01373	1.839	-0.02528	2.237
Standard deviation	0.1336		0.1186		0.1918	
Median absolute deviation	0.01762		0.01541		0.05687	
Mean absolute deviation	0.07034		0.0599		0.1211	

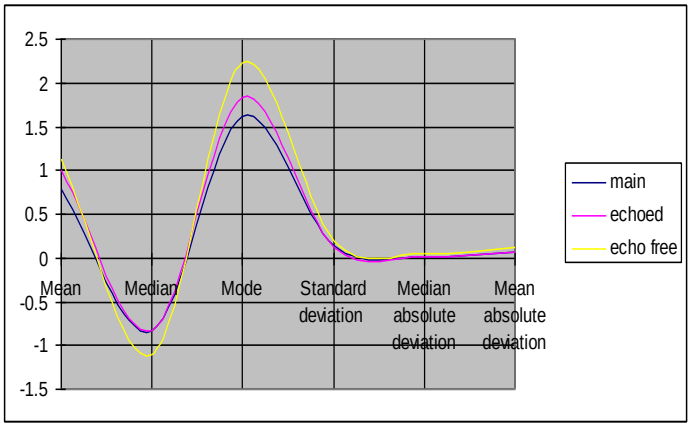


Figure 9: Comparing main signal, echoed signal and signal without echo.

Difference between the signal without echo and signal with echo is observed. The mean, median, mode all the values are analyzed more over the compressed signal is almost matched with the desire echo free signal , which is our goal.

IV. Conclusion

In this report, empirical audio signal has been considered to evaluate the performance of echo cancellation. It has been observed that the echo is suppressed without changing or distorting the main signal and user of VOIP can hear clear sounds. This program is easy to use and simpler than the conventional methods of suppressing echo. The simulation results have been tested and verified using Wavelet Tool. Power spectrum density has also been used to observe the difference of the received signal. The proposed system has found better performance in finding the echo cancellation than the conventional methods which can be adopted to suppress echo and take the fullest advantage of VOIP telephony.

References

[1] A Survey on Voice over IP over Wireless LANs Haniyeh Kazemitabar, Sameha Ahmed, Kashif Nisar, Abas B Said, Halabi B Hasbullah

[2] Echo in Voice over IP Systems Series VoIP Performance Management Date January 2006

[3] Implementation of an Acoustic Echo Canceller Using Matlab by Srinivasaprasath Raghavendran College of Engineering University of South Florida October 15, 2003

[4] IP Telephony: Packet-based multimedia communication system By Olivier Hersent, David Gurle, Jean-Pierre Petit

[5] Design a high-performance echo canceller for VOIP application, by Dr. Chang Y. Chob

[6] http://www.voipinfo.org/wiki/view/Asterisk+echo+cancellation

[7] http://www.eetimes.com/design/signal-processing-dsp/4017606/High-Performance-echo-canceller-for-Asterisk-VoIP-systems

[8] PACKET BASED ECHO CANCELLATION FOR VOICE OVER INTERNET PROTOCOL SIMULATED WITH VARIABLE AMOUNT OF NETWORK DELAY TIME by Ganesan Periakarruppan, Andy L.Y.Low, Hairul Azhar b Abdul Rashid

[9] LINE PROBING IN VOIP NETWORKS TO FIND PERFORMANCE LIMIT OF ECHO CANCELLER by Jerker Taudien, 2007

A New Test Method on the Convergence and Divergence for Infinite Integral

Guocheng Li
Linyi University at Feixian
Feixian, Shandong, P.R.China

***Abstract*—The way to distinguish convergence or divergence of an infinite integral on non-negative continuous function is the important and difficult question in the mathematical teaching all the time. Using the comparison of integrands to judge some exceptional infinite integrals are hard or even useless. In this paper, we establish the exponential integrating factor of negative function, and present a new method to test based on its exponential integrating factor. These conclusions are convenient and valid additions of the previously known results.**

Keywords-Iinfinite integral; Exponential integrating factor; Convergence and divergence.

I. INTRODUCTION

The convergence and divergence of infinite integral plays a significant role in mathematical analysis, and has been received much attention of many researchers. Many effective methods have been proposed such as Ordinary Comparison Test, Limiting Test, Dirichlet Test and Abel Test [1]. The basic ideas of these methods are to find another comparison infinite integral that its convergence or divergence is certain. But it is difficult or impossible to find the comparison infinite. So these arouse great interest of many scholars to research ([2-3]).

It is well-known that the convergence and divergence of infinite integral for the different integrand which its limit is zero is different. Base on the geometric meaning of infinite integral, it equal to the size of the area which is surrounded by the integrand (image) and X axis (not closed). The Convergence and divergence mainly determined by the proximity degree of integrand tends to the x-axis. In this paper, we discuss the test method for the convergence and divergence of infinite integral $\int_a^{+\infty} f(x)dx$ where $f(x)$ is the nonnegative continuous function which is defined in the interval $[a,+\infty), a \geq 0$ Although we have solved part of the problem by $\ln f(x)/\ln x$ in[3], it can't intuitive describe the proximity degree of integrand tends to the x-axis, so the exponential integrating factor of negative function is created and we obtain a new test method for the convergence and divergence of infinite integral. This method is more simple and feasible than ever before.

II. CHARACTERIZATIONS ON CONVERGENCE OF INFINITE INTEGRAL

Definition 2.1 Assume that $f(x)$ is non-negative differentiable function, we call the formula $R(x) = -f'(x)/f(x)$ is the function of the exponential integrating factor of $f(x)$, abbreviated as factor function.

From definition 2.1, we suppose the function $f(x) = f(x_0)\exp\left\{-\int_{x_0}^{x} R(t)dt\right\}, x_0 \geq a$ is completely determined by its factor function $R(x)$. When the independent variable tend to be infinite, we can use the size of factor function to compare the different functions reduce speed of which tend to the x-axis. For $\lim_{x\to+\infty} f(x) = 0$ is the necessary condition of convergence for infinite integrals $\int_a^{+\infty} f(x)dx$ ([1]), so we can easy to push that $R(x) \geq 0$ is the necessary conditions of convergence for $\int_a^{+\infty} f(x)dx$.

If $\lim_{x\to+\infty} \frac{f(x)}{g(x)} = c > 0$, we have $f(x)$ and $g(x)$ are equivalent. We denoted by $f \sim^t g$ .The nature of the equivalent functions is incredibly similar, so what about the infinite integration convergence of the equivalent integrand? We have the following conclusion.

Theorem 2.1 Assume that Non-negative differentiable functions $f(x)$ and $g(x)$ are equivalent, then its convergence and divergence is incredibly similar.

Proof: By $\lim_{x\to\infty} \frac{\int_a^x f(t)dt}{\int_a^x g(t)dt} = \lim_{x\to\infty} \frac{f(x)}{g(x)} = c > 0$,

There exist constant $m, M > 0$ such that

$m \le f(x)/g(x) \le M,\ \forall x > a.$ Obviously, we have

$$m\int_a^{+\infty} g(x)dx \le \int_a^{+\infty} f(x)dx \le M\int_a^{+\infty} g(x)dx;$$

$$\frac{1}{M}\int_a^{+\infty} f(x)dx \le \int_a^{+\infty} g(x)dx \le \frac{1}{m}\int_a^{+\infty} f(x)dx$$

From the definition of convergence of the infinite integrals, the desired result follows.

Theorem 2.2 Assume that $f(x)$, $g(x)$ are non-negative differentiable functions in interval $[a,+\infty)$ and $f(x)/g(x)$ is finally monotone decreasing, then

(i) $\exists x_0, \forall x > x_0, R_f(x) > R_g(x)$.

(ii) If $\int_a^{+\infty} g(x)dx$ is convergent, then $\int_a^{+\infty} f(x)dx$ is convergent, If $\int_a^{+\infty} g(x)dx$ is divergent, then $\int_a^{+\infty} f(x)dx$ is divergent.

Proof:

(i) $\exists x_0, \forall x > x_0$,

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{g(x)f'(x) - f(x)g'(x)}{g^2(x)} =$$

$$\frac{1}{f(x)g(x)}[R_g(x) - R_f(x)] < 0$$

$$\Leftrightarrow R_f(x) > R_g(x).$$

(ii) $\exists M > 0, let\ \frac{f(x)}{g(x)} \le M$, then $\int_a^{+\infty} f(x)dx \le M\int_a^{+\infty} g(x)dx$. Using Ordinary Comparison Test, we obtain this conclusion.

Assume that we know the convergence and divergence of an infinite integral, we can obtain the convergence and divergence of other infinite integrals by comparing the size of the exponential factor. But it's complex. In fact, it can be derived alone by the exponential factor $R(x)$ of integrand.

Theorem 2.3 Assume that $f(x)$ is non-negative differentiable function in $[a,+\infty)$,

(i) If $\lim_{x\to+\infty} xR(x) > 1$, then $\int_a^{+\infty} f(x)dx$ is convergent,

(ii) If $\lim_{x\to+\infty} xR(x) < 1$, then $\int_a^{+\infty} f(x)dx$ is divergent.

Proof: (i) $\forall k > 0$, we have $\frac{-f'(x)}{f(x)} \ge \frac{1+k}{x}$, then

$$-\ln f(x) \ge \ln(cx^{1+k}),\ c > 0.$$

That is $f(x) \le \frac{1}{c}x^{-(1+k)}$ i.e., According to the Ordinary Comparison Test, $\int_a^{+\infty} f(x)dx$ is convergent.

(ii) Let $g(x) = \frac{1}{x}$, then $\int_a^{\infty} \frac{1}{x}dx$ is divergent, for $R(x) < \frac{1}{x} = R_g(x)$, from Theorem 2.2, we have that $\int_a^{+\infty} f(x)dx$ is divergent.

Combining [3], we have that $\int_a^{+\infty} f(x)dx$ is possible convergent or divergent if $\lim_{x\to+\infty} xR(x) = 1$ from some possible examples. In fact, we can continue to study convergence as follow:

Theorem 2.4 Assume that $f(x)$ is non-negative differentiable function in interval $[a,+\infty)$ and

$$\lim_{x\to+\infty} xR(x) = 1.$$

(i) If $\lim_{x\to+\infty} \ln x[xR(x)-1] > 1$, then $\int_a^{+\infty} f(x)dx$ is convergent,

(ii) If $\lim_{x\to+\infty} \ln x[xR(x)-1] \le 1$, then $\int_a^{+\infty} f(x)dx$ is divergent.

Proof: (i) We can obtain that $R(x)$ is monotone-non-increasing function, for all $\forall k > 0$, we have

$$f(x) \le c_0 \exp\{-xR(x)\} \le$$

$$c_0 \exp\left\{-(1+\frac{1+k}{\ln x})\right\} <$$

$$c_0 \exp\left\{-\frac{1+k}{\ln x}\right\} = \frac{c_0}{x^{1+k}}$$

From Ordinary Comparison Test we can obtain that

$$\int_a^{+\infty} f(x)dx < +\infty$$

(ii)**Method1:** Let $R(x) \le \frac{1+\ln x}{x \ln x}$, we have that $-\ln f(x) \le \ln(x \ln x)$, i.e. $f(x) \ge \frac{1}{x \ln x}$, by $\int_a^{\infty} \frac{1}{x \ln x} dx$ is divergent, combining Ordinary Comparison test, we can obtain that $\int_a^{+\infty} f(x)dx$ is divergent.

Method2: Using the condition, we obtain that

$$R(x) \le \frac{1+\ln x}{x \ln x} = R_g(x),$$

Where $g(x) = \frac{1}{x \ln x}, and \int_a^{\infty} \frac{1}{x \ln x} dx$ is divergent. From theorem2.2, we have that $\int_a^{+\infty} f(x)dx$ is divergent too.

Using the conclusions above, we can easily think of the following examples.

Example 1: Discuss the convergence of $\int_2^{\infty} \frac{1}{x^p \ln^q x} dx$.

Solution: For its exponential integrating factor

$$R(x) = -\frac{p + cqx^c}{x},$$

We easy to know that $\lim_{x \to +\infty} x \frac{p \ln x + q}{x \ln x} = p$, using theorem 2.3, we have that the integration is convergent if $p > 1$, and is divergent if $p < 1$; when $p = 1$, we easy to know $x[R(x) - 1] = q$, using theorem 2.4, we get that the integration is convergent when $q > 1$; the integration is divergent when $q \le 1$.

Example2: Discuss the convergence of

$$\int_1^{\infty} x^p \exp(qx^c) dx.$$

Solution: For its exponential integrating factor is

$$R(x) = -\frac{p + cqx^c}{x},$$

We easy to know that $\lim_{x \to +\infty} xR(x) = -[p + cqx^c]$, using theorm2.3, we have $\int_1^{\infty} x^p \exp(qx^c) dx$ is convergence if $\lim_{x \to +\infty} p + cqx^c < -1$, i.e. $c \le 0$ or $q = 0, p < -1$ or c>0, q<0.

III. Constructing convergent Infinite Integral and possible approach

From Theorem 2.4, we can obtain that the convergence and divergence of $\int_a^{+\infty} f(x)dx$ is determined by the exponential integrating factor of $f(x)$.

Suppose that $\int_a^{+\infty} f(x)dx$ and $\int_a^{+\infty} f(x)g(x)dx$ are convergent, we give the following conditions of the non-negative function $g(x)$.

Using the exponential integrating factor of $f(x)g(x)$,

i.e. $R_{fg}(x) = R_f(x) + R_g(x)$,

if $\lim_{x \to +\infty} (x[R_f(x) + R_g(x)]) > 1$, then

$\int_a^{+\infty} f(x)g(x)dx$ is convergent. In fact, we can construct some convergent infinite integral for some specific needs.

Theorem 3.1 Assume that $\int_a^{+\infty} f(x)dx$ is convergent, if the non-negative function $g(x)$ is monotonous and bounded, then $\int_a^{+\infty} f(x)g(x)dx$ is convergent.

Proof: Combining the conditions and theorem 2.4, we know that $\lim_{x \to +\infty} \ln x[xR(x) - 1] > 1$. If $g(x)$ is decreasing and lower bounded,

Then $R_g(x) > 0$ and $R_{fg}(x) = R_f(x) + R_g(x)$ hold. Obviously, $\lim_{x \to +\infty} (x[R_{fg}(x) - 1]) > 1$; If $g(x)$ is increasing and upper bounded, then $\int_a^x R_g(t)dt$ is convergent, i.e. $\lim_{x \to +\infty} x \ln x R_g(x) = 0$, such that

$$\lim_{x \to +\infty} \ln x[xR_{fg}(x) - 1] = \lim_{x \to +\infty} \ln x[xR_f(x) - 1] > 1.$$

From all above, we can obtain that $\int_a^{+\infty} f(x)g(x)dx$ is convergent.

Theorem 3.2 Suppose $\int_a^{+\infty} f(x)dx$ is convergent where $f(x)$ is continuous and nonnegative in $[a, +\infty)$.

(i)If $\lambda < \infty$ and $\alpha < \lambda - 1$, then $\int_a^{\infty} x^{\alpha} f(x)dx$ is convergent.

(ii)If $\lambda = \infty, \forall \alpha \in R,$ then $\int_a^{\infty} x^{\alpha} f(x)dx$ is convergent,

(iii)If $\lambda = \infty$ and $\beta < \lim\limits_{x\to+\infty} \dfrac{\ln[\int_1^x [R(t)-\frac{1}{t}]dt+1]}{\ln x}$,

Then $\int_a^\infty \exp\{x^\beta\} f(x)dx$ is convergent, where

$$\lambda = \lim_{x\to+\infty} \inf\, xR(x),$$

Proof: (i) For $R(x)-\dfrac{\alpha}{x}$ is the factor function of $x^\alpha f(x)$, if $\alpha < \lambda - 1$, then $\lim\limits_{x\to+\infty} xR(x)-\alpha > 1$,i.e.

$$\lim_{x\to+\infty} x[R(x)-\frac{\alpha}{x}] > 1,$$

Using theorem 2.3, we get $\int_a^\infty x^\alpha f(x)dx$ is convergent.

(ii) Easy to get by (i).

(iii) From condition $x^\beta < \int_1^x [R(t)-\frac{1}{t}]dt+1$, i.e.

$$\int_1^x \beta t^{\beta-1}dt < \int_1^x [R(t)-\frac{1}{t}]dt,$$

We have $\beta x^{\beta-1} < R(x)-\dfrac{1}{x}$, from $\exp\{x^\beta\} f(x)$, we obtain its factor function $R(x)-\beta x^{\beta-1}$, using theorm2.3, we get $\lim\limits_{x\to+\infty} x[R(x)-\beta x^{\beta-1}] > 1$. Base on L'Hospital rule, we have

$$\beta < \lim_{x\to+\infty} \frac{xR(x)-1}{\int_1^x [R(t)-\frac{1}{t}]dt+1}.$$

Thus, we have that $\int_a^\infty \exp\{x^\beta\} f(x)dx$ is convergent.

IV. Conclusions And Prospect

In this paper, we get a new test method for the convergence and divergence of infinite integral, as everyone knows, defect integral can be changed into infinite integral by $x=\dfrac{1}{t}$ so we can expands our method on the defect integral.

We can almost judge the convergence of the infinite integrals by the conclusions refer to in Section 2. So we discuss some questions of infinite integral by the properties of the convergent infinite integral. In fact, the most important application in Section 3 is to construct plenty of distributions in the probability field especially in the Risk Theory. The claim distribution is vital for both the insured and the insurance company [4-7], constructing suitable distributions is beneficial to need for the insured or development of the insurance companies. This is a topic for future research.

Acknowledgment

The authors wish to give their sincere thanks to the anonymous referees for their valuable suggestions and helpful comments which improved the presentation of the paper.

References

[1] East China normal university, Mathematical analysis [M], Higher education press, 2001. (In Chinese)

[2] Wen Z.Y., Convergence of the infinite integrals of a new identification method [J], Mathematics,2005.21(2):111-112. (In Chinese)

[3] Luo L.P., The discussion of the necessary conditions for infinite integral convergence [J], Advanced mathematics research, 2005,8(4):19-21. (In Chinese)

[4] Embrechts P, Veraverbeke N. Estimates for the propabilty of ruin with special emphasis on the possibility of large claims. Insurancee Math Econ, 1982, 1: 55-72

[5] Klüppelberg C., Subexponential distributions and characterizations of relateed class. Probab. Th. Rel.Fields,1989, 82: 259-269.

[6] Cline D B H. Convolutions of distributions with exponential and subexponential tails. Austral Math Soc (Series A), 1987, 43: 347-365

[7] Yin Chuancun. A local theorem for the probability of ruin. Science in China (Series A). 2004, 47: 711-721.

Adaptive Outlier-tolerant Exponential Smoothing Prediction Algorithms with Applications to Predict the Temperature in Spacecraft

Hu Shaolin
State Key Laboratory of Astronautics
Xi'an, P.O.Box 505-16,710043,China

Li Ye
School of Automation,
Xi'an University of Technology, 710048, China

Zhang Wei
State Key Laboratory of Astronautics
Xi'an, P.O.Box 505-16,710043,China

Fan Shunxi
School of Automation,
Xi'an University of Technology, 710048, China

***Abstract*—The exponential smoothing prediction algorithm is widely used in spaceflight control and in process monitoring as well as in economical prediction. There are two key conundrums which are open: one is about the selective rule of the parameter in the exponential smoothing prediction, and the other is how to improve the bad influence of outliers on prediction. In this paper a new practical outlier-tolerant algorithm is built to select adaptively proper parameter, and the exponential smoothing prediction algorithm is modified to prevent any bad influence from outliers in sampling data. These two new algorithms are valid and effective to overcome the two open conundrums stated above. Simulation and practical results of sampling data from temperature sensors in a spacecraft show that this new adaptive outlier-tolerant exponential smoothing prediction algorithm has the power to eliminate bad infection of outliers on prediction of process state in future.**

Keywords-Exponential prediction; Adaptive smoothing prediction; Outlier-tolerance smoothing prediction.

I. INTRODUCTION

The exponential smoothing prediction algorithm was first suggested by Brown as an operation research analyst during World War II, when he worked on submarine locating and tracking model for antisubmarine warfare. He thought the trend of time series was stability or regularity, so the future state of time series could be deduced reasonably [1]. After more than half a century, exponential smoothing forecasting method has become a common forecasting method as well as the most commonly used forecasting method in short-term economic development trend forecast and many other areas in [1,2]. Exponential smoothing prediction method has obvious advantages: one is that the algorithm is simple, the predicted value at time t_{k+1} can be gained by actual measured value at time t_k and predicted value at time t_k, it is simple iterative relationship without complex operation; and the other is that it takes advantages from full-term average as well as moving average and doesn't abandon the old data. However, according to how long the sampling time of data away from the present moment, different weights are assigned to historical data and the influence of historical data on the current forecast gradually abates.

From the perspective of application, the most difficult point of exponential smoothing prediction algorithm is how to choose the parameter. On the one hand, the exponential smoothing prediction algorithm is very convenient for use because it need only one parameter; on the other hand, the accuracy and reliability of prediction results are changed when we set different value for the smooth parameter. In order to improve quality of prediction algorithm and prediction result, people did lots of explorations and researches from different angles and suggested many kinds of improved algorithms in the past half of the century. For example, some modification algorithms were suggested, such as the double, triple and multiple exponential smoothing prediction algorithms [1]. In recent years, people come up with a variety of adaptive algorithms and selection criteria [3-5] so as to find out some available selection method about smoothing parameter. In order to establish a new collaborative forecasting model, literature [6] also combined the exponential smoothing forecasting algorithm with neural network.

What is more, a lot of practice in using the exponential smoothing prediction show that the exponential smoothing prediction algorithm is lack of outlier-tolerant ability [7-8], because the exponential smoothing prediction with constant parameter is virtually linear prediction algorithm. In fact, the prediction results would inevitably produce serious distortion due to the impact of outliers when there are outliers in the sample sequences.

So, in order to improve the prediction quality and to make sure reliability of prediction results, there are two "bottleneck" problems which are open: one is how to select adaptively the smoothing parameter and the other is how to prevent bad

influence of outliers on the prediction algorithm. In this article we build firstly an useful online adaptive smoothing parameter estimation and then make further outlier-tolerant modification on prediction algorithm. At the end of this paper, a practical application is given to show that new algorithms are available.

II. ADAPTIVE FAULT-TOLERANT DESIGN OF PARAMETERS

The exponential smoothing prediction algorithm is based on the actual value $y(t_k)$ of the dynamic time-series in the t_k period and prediction value $\hat{y}(t_k)$ predicted at time t_k .The equations to calculate an exponential smoothing algorithm can be expressed as follows:

$$\begin{cases} \hat{y}(t_{k+1}) = (1-\alpha)y(t_k) + \alpha\hat{y}(t_k) \\ \hat{y}(t_1) = y(t_1) \end{cases} \quad (1)$$

where α is a smoothing constant parameter, $0 \le \alpha \le 1.$

How to select the parameter α is important because different values of α may bring on different predicting value of a dynamic process. For example, figure 1(a) is a series of simulation sampling data, the solid line and the dotted line in figure 1(b) are one-step prediction plots of figure 1(a) with two different parameter (α=0.2 and α=0.8) respectively. Comparing these two prediction curves shown in figure 1(b) clearly shown that the value of α has significant effect on smoothing prediction results.

How to choose the smoothing parameter α to get reliable prediction effect? An intuitive idea is to seek such a parameter estimator of α which satisfying the following relation:

$$S_k(\alpha) = \sum_{i=1}^{k} (y(t_i) - \hat{y}(t_i))^2 \xrightarrow{\alpha} \min \quad (2)$$

By the recursive formula (1), right side of the expression (2) can be written as follows

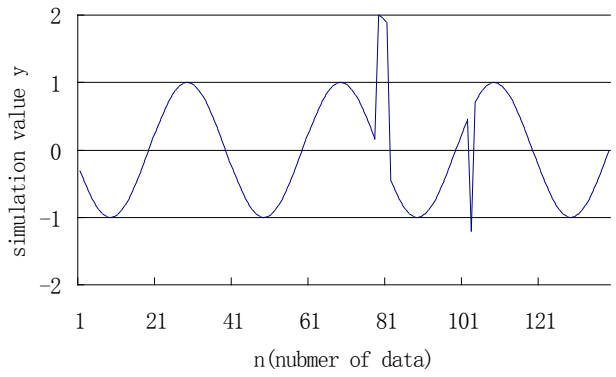
$$S_k(\alpha) = \sum_{i=1}^{k} \{[y(t_i) - y(t_{i-1})] + \alpha[y(t_{i-1}) - \hat{y}(t_{i-1})]\}^2$$

then the minimum point series of the parameter α in the function $S_k(\alpha)$ can be obtained as follows:

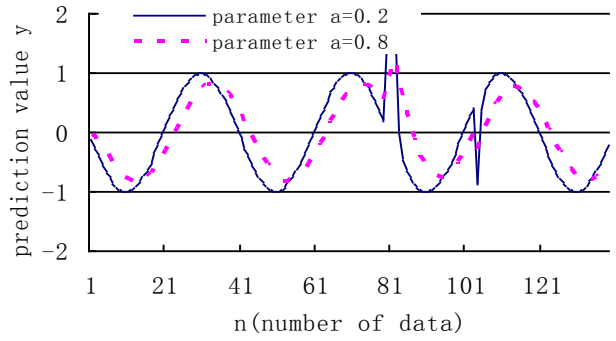
$$\hat{\alpha}_k^{(0)} = -\frac{\sum_{i=1}^{k}[y(t_i) - y(t_{i-1})][y(t_{i-1}) - \hat{y}(t_{i-1})]}{\sum_{i=1}^{k}[y(t_{i-1}) - \hat{y}(t_{i-1})]^2} \quad (3)$$

In order to ensure that design value of the smoothing parameter meets the constraint condition that the parameter α is a nonnegative real value and $0 \le \alpha \le 1$, we can change formula (3) into formula (4):

$$\hat{\alpha}_k = \max\{0, \min\{\hat{\alpha}_k^{(0)}, 1\}\} \quad (4)$$



(a) Smulation data plot



(b) Exponential smoothing prediction value plot

Figure 1. Simulation and exponential prediction plots

We can get online optimal estimation of smoothing coefficient from formula (3). Therefore, taking formula (1) and (3) into consideration, we can build a new form of adaptive exponential smoothing prediction algorithm as follows

$$\hat{y}(t_{k+1}) = y(t_k) + \frac{\sum_{i=1}^{k}[y(t_i) - y(t_{i-1})][y(t_{i-1}) - \hat{y}(t_{i-1})]}{\sum_{i=1}^{k}[y(t_{i-1}) - \hat{y}(t_{i-1})]^2}(y(t_k) - \hat{y}(t_k)) \quad (5)$$

with the beginning value $\hat{y}(t_1) = y(t_1)$.

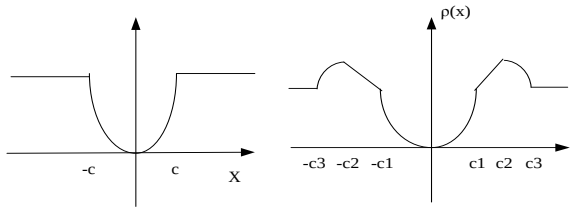
Using the formula (4), if quality of sampling sequence data is reliable, the ordinary exponential smoothing prediction algorithm (1) is changed into a new adaptive prediction algorithm (5), which can be used effectively to solve the open problem state in section I: how to choose the smooth coefficient and ensure that the forecast results can effectively track and early show the change of process.

The adaptive exponential smoothing prediction algorithm (5) is concise and practical. What's more, it is easy and convenient to choose the smoothing coefficient parameter. From engineering data processing and the actual process prediction viewpoint, the algorithm (5) as well as algorithm (1) does not have the ability of outlier-tolerance for isolated outliers and patchy outliers: when the sampling data series of a

dynamic process seriously deviate from the real state trend in a time or time period, subsequent predictions values of local arcs will distort or cause a big prediction deviation. So we will build a group of outlier-tolerant exponential smoothing prediction algorithms to improve outlier–tolerance ability of prediction algorithm for outliers and ensure that the predictions will not cause deviation even there are outliers occasionally in the sampling.

Following the famous research ideas about robust statistics suggested by Huber, we set the appropriate threshold parameter (c,c1,c2,c3) and select Huber-type $\rho-$ function [9] and the re-descending (short as Rd-)-type $\rho-$ function [10] in figure 2, then the formula (2) can be written as follows:

$$S_k(\alpha \mid \rho) = \sum_{i=1}^{k} \rho((y(t_i) - \hat{y}(t_i))^2) \xrightarrow{\alpha} \min \qquad (6)$$



(a) Huber-type $\rho-$ function (b) Rd-type $\rho-$ function

Figure 2. Two kinds of $\rho-$ function with fault-tolerance

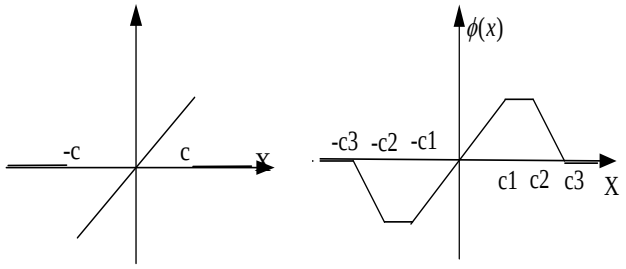
Similarly, we also can get an outlier–tolerant algorithm of for the exponential smoothing prediction parameter α:

$$\begin{cases} \hat{\alpha}_k^{(0)}(\phi) = -\dfrac{\sum_{i=1}^{k} \phi((y(t_i) - \hat{y}(t_i))^2)(y(t_i) - y(t_{i-1}))(y(t_{i-1}) - \hat{y}(t_{i-1}))}{\sum_{i=1}^{k} \phi((y(t_i) - \hat{y}(t_i))^2)(y(t_{i-1}) - \hat{y}(t_{i-1}))^2} \\ \hat{\alpha}_k(\phi) = \max\{0, \min\{\hat{\alpha}_k^{(0)}(\phi), 1\}\} \end{cases} \qquad (7)$$

Accordingly, algorithm (5) and algorithm (1) can be amended as the following form of adaptive fault-tolerant exponential smoothing prediction algorithm as follows:

$$\begin{cases} \hat{y}(t_{k+1}) = \hat{y}(t_k) + (1 - \hat{\alpha}_k(\phi))\phi(y(t_k) - \hat{y}(t_k)) \\ \hat{y}(t_1) = y(t_1) \end{cases} \qquad (8)$$

In formula (8), the $\phi-$ function $\phi(x) = \rho'(x)$. If the $\rho-$ function is the Huber-type function or Rd-type function in figure 2, then the $\phi-$ function is shown in figure 3.



(a) Huber type $\phi-$ function (b) Rd-type $\phi-$ function

Figure 3. Two kinds of $\phi-$ function with fault-tolerance

Intuitively, if the $\phi-$ function is from figure 3, then the exponential smoothing prediction algorithm (8) is outlier-tolerant. In fact, it can be seen from the figure 3 that so long as the forecast residual $\tilde{y}(t_k) = \tilde{y}(t_k) - \hat{y}(t_k)$ is below the reasonable scope threshold, two kinds of $\phi-$ function can make full use of innovation brought by new samples; However, once one-step predicting residual is beyond fixed threshold, the Huber-type $\phi-$ function has the ability to directly eliminate some negative influence from abnormal innovation on the of subsequent prediction; the Rd-type $\phi-$ function can make full use of the advantages of innovation and gradually reduce its impact on predicting results, according to the unusual degree with innovation.

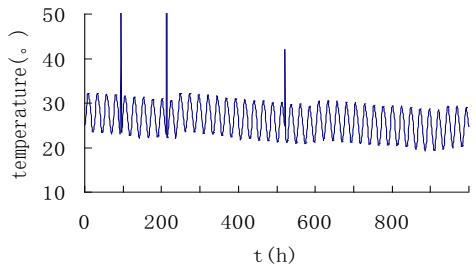
III. Actual Data Based Calculation And Analysis

Using data series as shown in figure 4(a) and the adaptive exponential smoothing prediction algorithm (5), the calculation results were shown in figure 4(b). Comparing the figure 4(b) with the figure 4(a), we may find out that the algorithm (5) can well forecast process variations. However, it is sensitive to outliers emerged in the sampling process, even causing a partial distortion to prediction results.

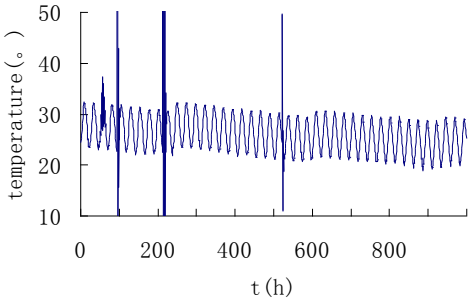
Using the formula (7)~(8) of the adaptive outlier-tolerant exponential smoothing prediction algorithm, and choosing the Rd-type $\phi-$ function $(c_1 = 3\sigma, c_1 = 5\sigma, c_1 = 7\sigma)$,we get one-step ahead prediction plot shown in figure 4(c). As can be clearly seen from these calculation data sequence, using the adaptive fault-tolerant exponential smoothing prediction algorithm can effectively avoid the adverse impact on outliers in sampling sequence and accurately predict the change of process status.

IV. Conclusion

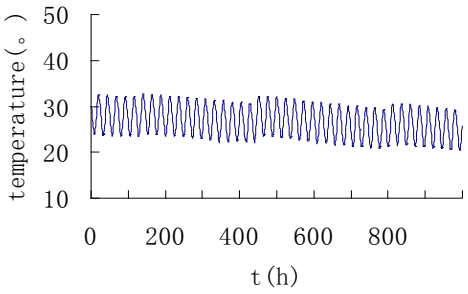
Comparing figure 4(c) with figure 4(a),it is shown that the adaptive outlier-tolerant exponential smoothing prediction algorithm has the ability to forecast process variations accurately under normal conditions for data sampling. Even if there are a few outliers in sampling data, the adaptive outlier-tolerant exponential smoothing prediction algorithm can also do well.



(a) Actual data plot



(b) Adaptive exponential smoothing prediction plot



(c) Adaptive fault-tolerant exponential smoothing prediction plot

Figure 4. Adaptive exponential smoothing prediction and adaptive fault-tolera nt exponential smoothing prediction curves of spacecraft sensor temperature

The research contents of this paper and the results obtained have widely application value. The adaptive outlier-tolerant exponential smoothing prediction algorithm can be widely used in many different fields, such as space control and process monitoring and economic forecasting and sensor fault detection.

In fact, a lot of change in dynamic process can be transformed into abnormal changes in measurement data or in system state.

ACKNOWLEDGMENT

The research was supported by the National Nature Science Fund of China (No.61074077) and the Tianyuan Fund of National Nature Fund of China (No.11026224).

REFERENCES

[1] Gardner ES Jr. 1985. Exponential smoothing: the state of the art. Journal of Forecasting 4: 1-28, 1985.

[2] Xiangbao Gao, Hanqing Dong. Data analysis and application of SPSS, Beijing: Tsinghua University Press, 2009.4.

[3] R Snyder and J Ord. Exponential smoothing and the Akaike information criterion. Monash University,2009.6

[4] S Makridakis,S Wheelwright and R Hyndman. Forecasting: methods and application. New York, John Wiley & Sons. 1998

[5] Changjiang Wang. Research on smooth coefficient choice with exponential smoothing method. Journal of North University of China, vol .6:6-12,2006

[6] Kin Keung Lai, Lean Yu and Shouyang Wang. Hybridizing exponential smoothing and neural network for Financial Time Series Predication. V.N. Alexandrov et al. (Eds.): ICCS 2006, Part IV, LNCS 3994, pp. 493 – 500, Springer-Verlag Berlin Heidelberg,2006.

[7] Hu Shaolin, Sun Guoji. Process monitoring technology and application (in Chinese).Beijing:National Defence Industry Press ,2001

[8] Hu Shaolin,Wang Xiaofeng,Karl Meinke,Huajiang Ouyang.Outlier-tolerant fitting and online diagnosis of outliers in dynamic process sampling data series.in Artificial Intelligence andComputational Intelligence, pp.195-204, LNAI 7004(Deng Miao,Lei Wang,eds) Springer Press,2011

[9] P J Huber. Robust statistics,John &Sons Press,1981

[10] Fan Jincheng,Hu Feng. Minimax robustly redescending estimators for multivariate location vector.Journal of Xi'an Jiaotong University, vol. 29,no.12,pp:107-112,1995

AUTHORS' PROFILE

Dr. Hu Shaolin graduated from Xi'an Jiaotong University of China and received his PhD degree in Control Science and Engineering in 2000. He pursued his postdoctoral studies in University of Science and Technology of China in 2005. Currently, Dr. Hu Shaolin is a professor and researcher at the State Key Laboratory of Astronautics. His research interests and publications are in Control Engineering, System Safety, Digital Signal Processing, and Data Mining.

Miss Li Ye graduated from Zhongbei University of China and received her bachelor's in Automation in 2010. Currently, Miss Liye is a postgraduated student at Xi'an University of Technology of China. Her research interests are in navigation and digital signal processing.

Mr Zhang Wei graduated from Southeast University of China and received her master's degree in Automation in 2006. Currently, Mr Zhangwei is an engineer at the State Key Laboratory of Astronautics..His research interests are in digital signal processing.

Mr Fan Shunxi graduated from Henan Normal University of China and received his bachelor's degree in Automation in 2009. Currently, Mr Fan is a postgraduated student at Xi'an University of Technology of China. His research interests are in outlier-tolerant Kalman filter with applications in navigation and digital signal processing.

Towards Quranic reader controlled by speech

Yacine Yekache, Yekhlef Mekelleche, Belkacem Kouninef
Institut National des Télécommunications et des TIC
BP 1518 Oran El Mnaouer 31000 Oran, ALGERIA

***Abstract*—In this paper we describe the process of designing a task-oriented continuous speech recognition system for Arabic, based on CMU Sphinx4, to be used in the voice interface of Quranic reader.**
The concept of the Quranic reader controlled by speech is presented, the collection of the corpus and creation of acoustic model are described in detail taking into account a specificities of Arabic language and the desired application.

Keywords-arabic speech recognition; quranic reader; speech corpus; HMM; acoustic model.

I. INTRODUCTION

Automatic speech recognition (ASR) is the technology that permits the communication with machine using speech. There are several applications that use this technology such as hands-free operation and control as in cars or person with disabilities, automatic dictation, government information systems, automatic query answering, telephone communication with information systems, etc.

The most dominant approach for ASR system is the statistical approach Hidden Markov Model(HMM), trained on corpora that contain speech resource from a large number of speakers to achieve acceptable performance; unfortunately there is a lack of this corpus for Arabic language. In this work we collected a new corpus called Quranic reader command and control which we will use to create an acoustic model using "sphinx train"

II. QURANIC READER AND ARABIC LANGUAGE

A. Quranic reader

Quran is the central religious text of Islam, which is the verbatim word of God and the Final Testament, following the Old and New Testaments. It is regarded widely as the finest piece of literature in the Arabic language. The Quran consists of 114 chapters of varying lengths, each known as a sura. Chapters are classed as Meccan or Medinan, depending on when (before or after Hijra) the verses were revealed. Chapter titles are derived from a name or quality discussed in the text, or from the first letters or words of the sura.

There is a crosscutting division into 30 parts of roughly equal division, ajza, each containing two units called ahzab, each of which is divided into four parts (rub 'al-ahzab).

The Quran is the muslims way of life and the guidance from Allah for that every muslim should read, listen and memorize it; nowadays there are computer tools used for this purpose, the interaction with this tools is by using a mouse or a keyboard but in some situation it is difficult to use them for example when driving a car or for blind person; so our goal is to create a Quranic reader controlled by speech

B. Arabic language

Quran is revealed in Arabic for that it is the official language of 23 countries and has many different, geographically distributed spoken varieties, some of which are mutually unintelligible. Modern Standard Arabic (MSA) is widely taught in schools, universities, and used in workplaces, government and the media.

Standard Arabic has basically 34 phonemes, of which six are vowels, and 28 are consonants. A phoneme is the smallest element of speech units that makes a difference in the meaning of a word, or a sentence. The correspondence between writing and pronunciation in MSA falls somewhere between that of languages such as Spanish and Finnish, which have an almost one-to-one mapping between letters and sounds ,and languages such as English and French, which exhibit a more complex letter-to-sound mapping[1] .

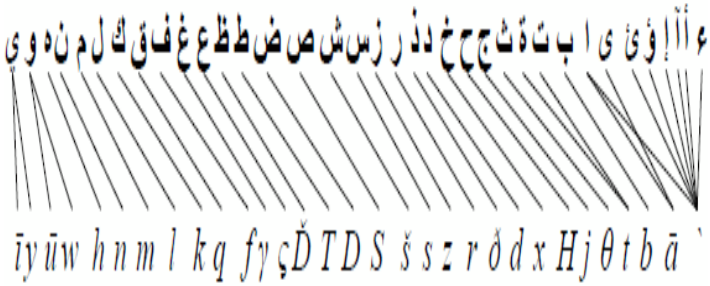


Figure 1. letter to sound mapping in Arabic

$$\widehat{W} = \underset{W}{argmax}\, P\left(W/X\right) \quad (1)$$

III. STATISTICAL APPROACH FOR SPEECH RECOGNITION

A. Mathematical formulation of ASR problem

A block diagram of this fundamental approach to speech recognition is given in Fig.2, which shows a sentence W being converted to a speech signal s[n] via the speech production process. The speech signal is then spectrally analyzed (by the acoustic processor) giving the spectral representation, X = (X1, X2, . . . , XL) for the L frames of the speech signal. The linguistic decoder solves for the maximum likelihood sentence that best matches X (i. e., maximizes the probability of W given X) via the Bayesian formulation:

$$\widehat{W} = \underset{W}{argmax}\ \frac{P(W/X)P(W)}{P(X)} \quad (2)$$

$$\widehat{W} = \underset{W}{argmax}\ P_A(X/W)\,P_L(W) \quad (3)$$

Step 3 Step 1 Step 2

The maximization of (1) is converted to (2) using the Bayes rule, and since the denominator term P(X) is independent of W, it can be removed leading to the three-step solution of (3). Here we explicitly denote the acoustic model by labeling P(X|W) as PA(X|W), where A denotes the set of acoustic models of the speech units used in the recognizer, and we denote P(W) as PL (W) for the language model describing the probabilities of various word combinations. The process of determining the maximum-likelihood solution is to first train (offline) the set of acoustic models so that step 1 in (3) can be evaluated for each speech utterance. The next step is to train the language model for step 2, so that the probability of every word sequence that forms a valid sentence in the language model can be evaluated. Finally step 3 is the heart of the computation, namely a search through all possible combinations of words in the language model to determine the word [2]

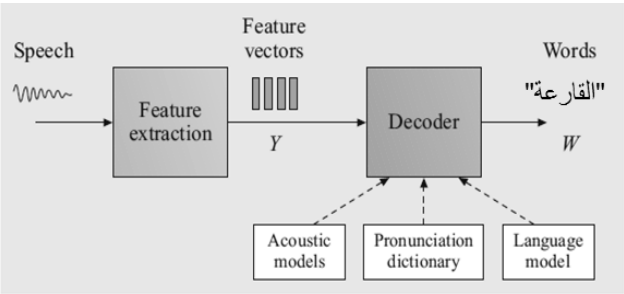


Figure 2. Architecture of an HMM based recognizer

B. CMU Sphinx

Sphinx4 is a software implementation of HMM speech recognizer, it's architecture is highly flexible, Each labelled element in Figure (3) represents a module that can be easily replaced, allowing researchers to experiment with different module implementations without needing to modify other portions of the system. The main blocks in Sphinx-4 architecture are frontend, decoder and Linguist [3]

Front End: it parameterizes an input signal into a sequence of output features. It performs Digital Signal Processing on the incoming data.

--Feature: The outputs of the front end are features, used for decoding in the rest of the system.

Linguist: Or knowledge base, it provides the information the decoder needs to do its job, this sub-system is where most of the adjustments will be made in order to support Arabic recognition it is made up of three modules which are:

--Acoustic Model: Contains a representation of a sound, created by training using many acoustic data.

--Dictionary: It is responsible for determining how a word is pronounced.

--Language Model: It contains a representation (often statistical) of the probability of occurrence of words.

Search Graph: The graph structure produced by the linguist according to certain criteria (e.g., the grammar), using knowledge from the dictionary, the acoustic model, and the language model.

Decoder: It reads features from the front end, couples this with data from the knowledge base and feedback from the application, and performs a search to determine the most likely sequences of words that could be represented by a series of features.

The Configuration Manager: This sub-system is designed in order to make Sphinx-4 pluggable and flexible. Configuration information is provided in XML files, and the actual code uses the Configuration Manager to look up the different modules and sub-systems.

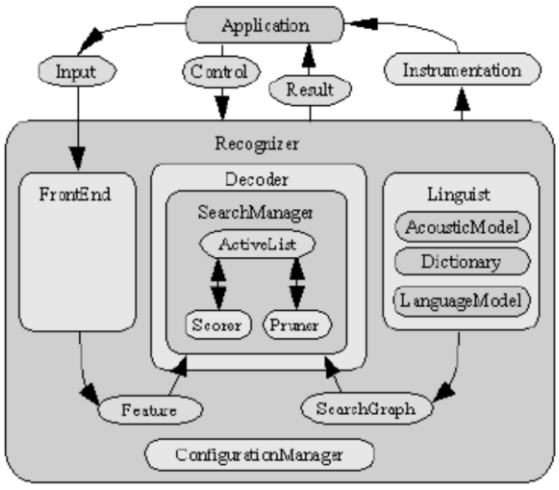


Figure 3. Sphinx-4 architecture

The Sphinx framework was originally designed for English, but nowadays it supports also, among others, Spanish, French and Mandarin. However, there are limited acoustic model available for Arabic [4, 5, 6].

Our goal is to use Sphinx to build Quranic reader controlled by speech ; the main tuning that we should do in the frame work is the creation of the acoustic model and the application that interact with the framework. In the following section we will describe the creation of the acoustic model.

IV. ACOUSTIC MODELING

The creation of acoustic model pass through two steps, the first is data collection (corpus) and the second is training the model using this data,

A. Data collection (corpus)

The following elements were required to train a new acoustic model:

- Audio data with recorded speech;
- Transcription of each audio file;
- Dictionary with phonetic representations of all words appearing in the transcriptions;
- List of phonemes (and sounds) appearing in the transcriptions

B. Speech collection

We prepared a text file which contain 114 suras name's, famous receiters names and control words that must be correctly recognized for the Quranic reader to function properly, The amount of audio data required to properly train the model depends on the type of the model. For a simple command-and-control one-speaker system the amount of data can be fairly low. For multi-speaker systems the amount of required audio increases and this is our case.

After selecting text for the recognizer, recording of this chosen data is required. For this work, the recording has been taken place using 50 speakers from different region of Algeria; their personal profile includes information like name, gender, age, and other information like Environmental condition of recording (for example: class room condition, sources of noise like fan, generator sound etc) Technical details of device (pc, microphone specification)

The audio file was recorded using sampling rate of 16KHZ and 16 bit per sample, this rate was chosen because it provide more accurate high frequency information, after that the splitting by command and word was done manually and saved in .wav format[7]. Each file has been named using this convention: speakername-commandID.wav for example a file with the name yacine-001 mean that this file is recorded by yacine and it contain the Fatiha word

These audio files was divided into two sets, the first is composed of data from 20 males and 15 female used to train the acoustic model and the second composed of data from 10 males and 5 female for testing purpose.

C. Transcription file

The second step is the transcription of the training set of the collected audio files; any error in the transcription will mislead the training process later. The transcription process is done manually, that is, we listen to the recording then we match exactly what we hear into text even the silence or the noise should be represented in the transcription.

D. Phonetic dictionary

In this step we mapped each word in the vocabulary to a sequence of sound units representing pronunciation; that it contained all words with all possible variants of their pronunciation, to take into account pronunciation variability, caused by various speaking manners and the specificity of Arabic. . Careful preparation of phonetic dictionary prevents from incorrect association of a phoneme with audio parameters of a different phoneme which would effect in decreasing the model's accuracy.

For example

صِفْر ص ِف ر

خمسة خَ م س ه

النَّاس أ نْ ن ا س

E. List of phoneme

This is a file which contain all the acoustic units that we want to train model for, The SPHINX-4 does not permit us to have units other than those in our dictionaries. All units in the dictionary must be listed here. In other words, phone list must have exactly the same units used in your dictionaries, no more and no less. The file has one phone in each line, no duplicity is allowed.

TABLE1. WORDS USED IN THE CORPUS

Sura name	الفاتحة البقرة آل عمران النساء	001 002 003 004 114
Reciters names	الغامدي السديسي العجمي الحذيفي	115 116 117 118
Control	تلاوة إنهاء توقف إستمر تكرار تحفيظ تفسير إنتقل بحث سورة آية حزب جزء نفد	119 120 121 122 123 124 125 126 127 128 129 130 131 132
Arabic digit	صفر واحد .. تسعة	133 134 142

F. Acoustic model training

Before acoustic modeling we should extract features vectors from the speech for a purpose of training. The

dominating feature extraction technique known as Mel-Frequency Cepstral Coefficients (MFCC) was applied to extract features from the set of spoken utterances. A feature vector Y represents unique characteristics of each recorded utterance,

The most widely used method of building acoustic models is HMMs, Each base phone q is represented by a continuous density HMM of the form illustrated in Fig.(4) with transition parameters {aij } and output observation distributions {bj ()}.entry and exit states are nonemitting and they are included to simplify the process of concatenating phone models to make words.

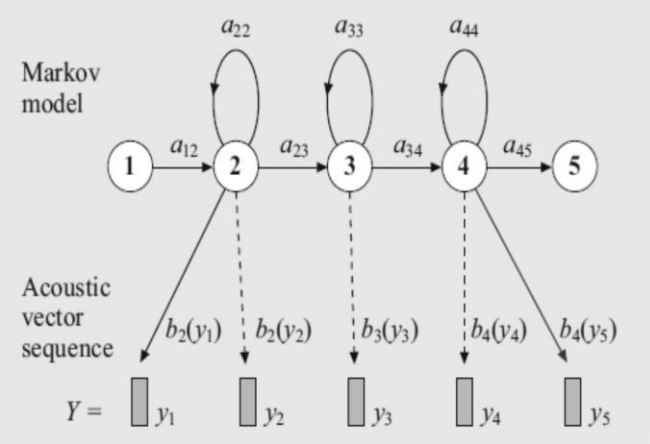


Figure 4. HMM based phone model

Each word in our corpus should be modeled using HMM,the parameter aij and bj are estimated from our collected corpus using expectation maximization (EM). For each utterance, the sequence of base forms is found and the corresponding composite HMM constructed.

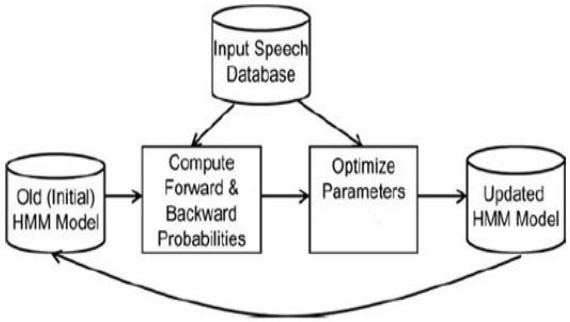


Figure 5. The Baum–Welch training procedure based on a given training set of utterances

A forward–backward alignment is used to compute state occupation probabilities (the E step) and the means and covariances are then estimated via simple weighted averages(theMstep) [8]

To create the acoustic model we used sphinx train, which need as input the recorded speech, transcription, dictionary and phoneme files to produce the acoustic model. Much of Sphinx Train's configuration and setup uses scripts written in the programming language Perl.

V. CONCLUSION

In this paper we reported the first steps toward developing Quranic reader controlled by speech using sphinx4 framework, in this steps we specified the words that should be recognized and collected a corpus used to train the acoustic model with sphinx train.

Further we will integrate the acoustic model in sphinx 4 and build an application that interacts with sphinx framework.

REFERENCES

[1] Nizar Y. Habash, "Introduction to arabic natural language processing" 2010 by Morgan & Claypool.

[2] J. L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition",Proc. IEEE 77(2), 257–286 (1989).

[3] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, Joe Woelfel, "Sphinx-4: A Flexible Open Source Framework for Speech Recognition,". SMLI TR2004-0811 c2004 SUN MICROSYSTEMS INC

[4] H. Hyassat, and R. Abu Zitar, "Arabic speech recognitionusing SPHINX engine," International Journal of SpeechTechnology, Springer, pp. 133-150,2008.

[5] K. Kirchhoff, J. Bilmes, S. Das, et aI., "Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins summer workshop," ICASSP'03, Hong Kong, vol.l, pp. 344-347,2003

[6] M. Ali, M. Elshafei, M. Alghamdi, H. Almuhtaseb, and A. Al-Najjar, "Generation of Arabic Phonetic Dictionaries for Speech Recognition," IEEE Proceedings of the International Conference on Innovations in Information Technology, UAE, pp. 59 – 63, 2008.

[7] Artur Janicki , Dariusz Wawer "Automatic Speech Recognition for Polish In a Computer Game Interface" Proceedings of the Federated Conference on Computer Science and Information Systems pp. 711–716

[8] Jacob Benesty, M. Mohan Sondhi, Yiteng Huang "Springer handbook of speech processing" partE p 554

[9] Alghamdi, Mansour, Fayez Alhargan, Mohamed Alkanhal,Ashraf Alkhairy, Munir Eldesouki and Ammar Alenazi1 (2008) Saudi Accented Arabic Voice Bank. Workshop on.Experimental Linguistics. Athens, Greece

[10] Artur Janicki , Dariusz Wawer "Automatic Speech Recognition for Polish In a Computer Game Interface" Proceedings of the Federated Conference on Computer Science and Information Systems pp. 711–716

[11] Jacob Benesty, M. Mohan Sondhi, Yiteng Huang "Springer handbook of speech processing" partE p 554.

[12] "Introduction to Arabic Speech Recognition Using CMUSphinx System" by H Satori, M Harti, N Chenfour

A novel approach for pre-processing of face detection system based on HSV color space and IWPT

Megha Gupta, Prof. Neetesh Gupta
Information and Technology department
Technocrats Institute of Technology
Bhopal, Madhya Pradesh, India

***Abstract*— Face detection system is challenging area of research in the field of security surveillance. Preprocessing of facial image data is very important part of face detection system. Now days various method of facial image data preprocessing are available, but all these method are not up to mark. Now we have a novel approach for preprocessing of face detection system based on HSV color space and integer wavelet packet transform method. And finally in this paper we have used LBP (local binary pattern) and SVM classification process are used.**

Keywords- facial image, HSV color space, IWPT, SVM.

I. INTRODUCTION

In today's world human face detection is a prelude required step of face recognition systems as well as very important process in security based applications. Human often uses faces to recognize persons and advancement in computing capability over the past few decades. The development of face detection system is quite essential in a variety of application such as robotics, security system, and intelligent human-computer interfaces, etc. A number of face detection methods such as those using Eigen-faces [1] and neural network [2], have been developed. In these methods however a large amount of computation is required, making the processing extremely time consuming. Initial step of any processing of face is detecting the location in images where faces are present. But face diction from a single image is difficult because of variability in scale, locations, pose, and color.

There are many methods for the detection of face; color is an important feature of human face. Using skin color as feature for extracting a face increases the detection rate. Face recognition can be used for both verification and identification. Verification is done by comparing the two images. If the two images got matched means the input image is verified. Identification is done by comparing the input image with more than one image and finding the closest match of that input image. The method presented in this paper consists of three steps- skin reorganization, face reorganization and face recognition. The innovation of the proposed method is using skin reorganization filter as a preprocessing step for face reorganization. Now in this paper we proposed new preprocessing technique using integer wavelet packet transform and HSV color space for better preprocessing of facial image data. Integer wavelet packet transform function is a special function of transform function from these function we decompose the facial image data without loss of information. And approach of this paper is HSV color space for better intensity of facial image instead of RGB model. The remainder of this paper is organized as follows. Section II briefly reviews the literature. A new approach for preprocessing of facial image data is explained in Sect. III. Finally, section IV presents conclusions and future scope

II. LITERATURE REVIEW

By Mayank Vatsa, Richa Singh [1] proposed method for facial image preprocessing as Age Transformation using Mutual Information Registration the face images are pre-processed and quality of the image is normalized, we minimize the age difference between gallery and probe face images. One way to address the challenge of facial aging is too regularly update the database with recent images or templates. However, this method is not feasible for applications such as border control and homeland security, missing persons and criminal investigations address this challenge, researchers have proposed several age simulation and modeling techniques. These technique model the facial growth that occurs over period of time to minimize the difference between probe and gallery images. Unlike, the conventional modeling approach, we proposed mutual information registration based age transformation algorithm to minimize the age difference between gallery and probe images. Mutual information is a concept from information theory in which statistical dependence is measured between two random variables [3, 4].

Researchers in medical imaging have used mutual information based registration algorithms to effectively fuse images from different modalities such as CT and MRI [3], [4]. We have used this algorithm because there may be variations in the quality of pre-processed scanned and digital face images, and the registration algorithm should contend with these variations. Age difference minimization using registration of gallery and probe face images is described as follows:

Let F_G and F_P be the detected and quality enhanced gallery and probe face images to be matched. Mutual information between two face images can be represented as,

$$M(F_G, F_P) = H(F_G) + H(F_P) - H(F_G, F_P) \quad (1)$$

Where, H (.) is the entropy of the image and H (F_G, F_P) is the joint entropy. Registering F_G with respect to F_P requires maximization of mutual information between F_G and F_P, thus maximizing the entropy H (F_G) and H (F_P), and minimizing the joint entropy H (F_G, F_P). Mutual information based registration algorithms are sensitive to changes that occur in the distributions as a result of difference in overlap regions. To address this issue, Hill et al. [3] proposed normalized mutual information that can be represented as,

$$N\,M\,(F_G, F_P) = \frac{H(F_G)+H(F_P)}{H(F_G, F_P)} \qquad (2)$$

The registration is performed on the transformation space, T, such that

$$T = \begin{pmatrix} a & b & 0 \\ c & d & 0 \\ e & f & 1 \end{pmatrix}$$

Where a, b, c, d are shear, scale, and rotation parameters, and e, f are the translation parameters. Using the normalized mutual information and exploring the search space, T, we define a search strategy to find the transformation parameters, T *.

$$T^* = \arg\max \{N\,M\,(F_P, T\,(F_G))\} \qquad (3)$$

Gallery and probe face images (F_G and F_P) are thus registered using the transformation T *. This registered algorithm is linear in nature. To accommodate nonlinear variations in faces, multi resolution image pyramid scheme is applied which starts with building the Gaussian pyramid of both the gallery and probe images. Registration parameters are estimated at the coarsest level and used to warp the face images in the next level of the pyramid. The process is iteratively repeated through each level of the pyramid and a final transformed gallery face image is obtained at the finest pyramid level.

In this manner, the global variations at the coarsest resolution level and local non-linear variations the finest resolution level are addressed. Once the age difference between the gallery and probe face image is minimized, face recognition algorithm can be efficiently applied to verify the identity of the probe image, By Hazym Kemal Ekenel, Hua GAO[3]proposed method for facial image preprocessing Discrete Cosine Transform-Based Local Appearance Model .

Local appearance face recognition is bases on statistical representations of the non- overlapping local facial regions and their combination at the feature level. The underlying idea is to utilize local information while preserving spatial relationships. In[5], the discrete cosine transform (DCT) is proposed to be used to represent the local regions. It has been shown to be a better representation method for modeling the local facial appearance compared to principal component analysis(PCA) and the discrete wavelet transform (DWT) in terms of face recognition performance. Feature extraction from depth images using local appearance-based face representation can be summarized as follows:

The input depth image is divided into blocks of 8x8 pixels size. Each block is then represented by its DCT coefficients. These DCT coefficients are ordered using the zigzag scanning pattern [6]. From the ordered coefficients, M of them is selected according to the feature selection strategy, resulting in an M-dimensional local feature vector. Finally, the DCT coefficients extracted from each block are concatenated to construct the overall feature vector of the corresponding depth image. In order to compare the introduced local DCT-based representation with the depth representation, we calculated the ratio of within class variance with each representation on a training set which has also been used for identification experiments.

We calculated the ratio of within class variance to between class variance for each representation unit and then averaged it over the representation units and the subjects. We obtained an average ratio of 0.5 with DCT-based local representation, and 0.67 with the depth representation. The lower ratio of within class variance to between class variance obtained by the proposed representation scheme indicates its better discrimination capability compared to the depth representation

III. A New Approach for Facial Image Preprocessing

In this section we have discuss the transform of facial image data and resolve the color of skin into HSV color space and local binary pattern method. First we have discuss integer wavelet packet transform function.

A. Integer Wavelet Packets Transform

In this section the wavelet transform and of the filter bank scheme are given, and the wavelet packets transform are introduced. The block scheme of the single level wavelet transform is shown in Fig.1.

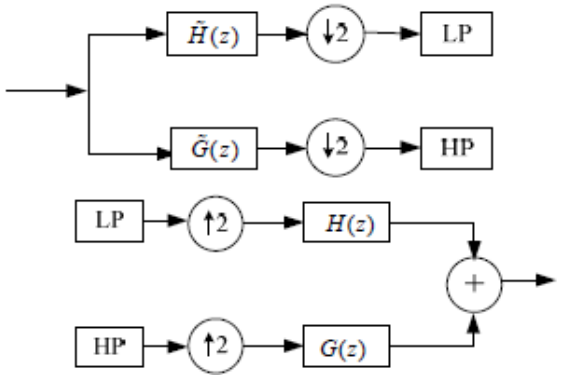


Fig. (1) One level transforms function

The low-pass analysis filters and the high-pass ones are followed by down sampling of a factor two. At the reconstruction side, the low-pass and band-pass branches are psampled and filtered with the synthesis filters $H(z)$ and $G(z)$ in order to obtain the original signal A wavelet transform on J levels is obtained by iterating the filter bank J-1 times on the low-pass branch. The wavelet transform coefficients consist of the J high-pass and the erminal low-pass node sequences output by the filter bank tree. Given a perfect reconstruction filter bank, the iterated scheme represents an either orthonormal or biorthogonal (non-redundant) representation of the original

signal. Differently from the wavelet transform, the J-level WPT are achieved by iterating the one level filter bank on both the low-pass and the high-pass branch, and then applying a pruning algorithm to select a suitable representation. An algorithm has been proposed in [5], which selects the best representation of a sequence across the entire tree based on some proper cost function, which must measure the compactness of the representation.

B. HSV colour model

The HSV stands for the Hue, Saturation, and Value based on the artists (Tint, Shade, and Tone). The coordinates system in a hexacone is shown in Figure 1. (a) And Figure 1. (b) a view of the HSV color model. The Value represents intensity of a color, which is decoupled from the color information in the represented image. The hue and saturation components are intimately related to the way human eye perceives color resulting in image processing algorithms with physiological basis

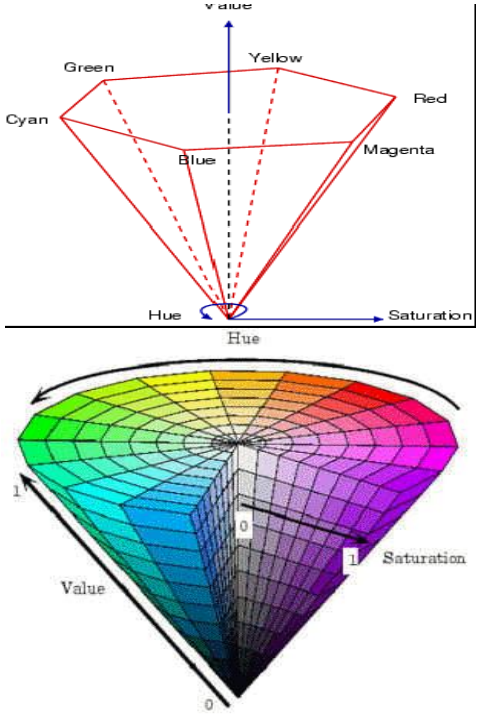


Fig. (2) HSV color model

C. SVM classifier

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A schematic example is shown in the illustration below. In this example, the objects belong either to class GREEN or RED.

The separating line defines a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED. Any new object (white circle) falling to the right is labeled, i.e., classified, as GREEN (or classified as RED should it fall to the left of the separating line).

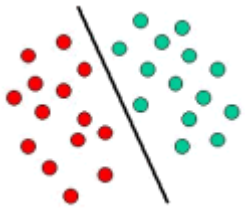


Fig. (3) Classification of data through SVM

The above is a classic example of a linear classifier, i.e., a classifier that separates a set of objects into their respective groups (GREEN and RED in this case) with a line. Most classification tasks, however, are not that simple, and often more complex structures are needed in order to make an optimal separation, i.e., correctly classify new objects (test cases) on the basis of the examples that are available (train cases). This situation is depicted in the illustration below. Compared to the previous schematic, it is clear that a full separation of the GREEN and RED objects would require a curve (which is more complex than a line.

Steps for processing of preprocessing of facial image data

(1) Processed scanned image of face into transform function.
(2) Apply integer wavelet packet transform function for decomposition of image on the basis of low and high frequency without loss of data.
(3) Built packet of image data.
(4) Apply HSV color model for better resolution for facial skin.
(5) Finally used SVM clasifer for classification of low data and high data
(6) Then apply LBP (local pattern method)

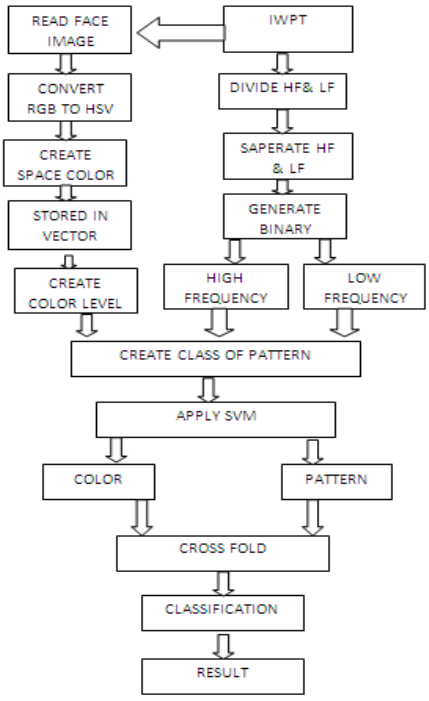


Fig.(4) Flow chart

IV. Conclusion and Future Scope

In this paper we discuss a combined pre-processing method for facial image data for better processing of raw data for training and detection of face. We have seen here by mathematics integer wavelet packet transform is a lossless decomposition technique and very efficient process for image transformation. Now in future we have implanted this approach and find some result and these results compare with standard methods

References

[1] Mayank Vatsa, Richa Singh, Samarth Bharadwaj, Himanshu S. Bhatt and Afzel Noore" Matching Digital and Scanned Face Images with Age Variation" IEEE Transactions on Circuits, Systems and Video Technology, vol. 14, no. 1, pp. 50–57, 2010

[2] Liu, C. and Wechsler, H. (2002). "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition" Image processing, IEEE Transactions on 11:467-476.

[3] Romby, H.A, Baluja, S. and Kanade T. (1998). "Neural network based face detection". Pattern Analysis and Machine Intelligence, IEEE Transactions on 20:23-38

[4] D. Hill, C. Studholme, and D. Hawkes, "Voxel similarity measures for automated image registration", in Proceedings of Third SPIE Conference on Visualization in Biomedical Computing, 1994, pp.205-206.

[5] F. Maesa, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information",IEEE Transactions on Medical Imaging, vol. 6, no. 2,pp.187–198, 1997.

[6] H. K. Ekenel and R. Stiefelhagen, "Local appearance-based face recognition using discrete cosine transform," presented at the 13th Eur. Signal Processing Conf., Antalya, Turkey, 2005.

[7] R. C. Gonzales and R. E. Woods, *Digital Image Processing*. Upper Saddle River, NJ: Prentice-Hall, 2001

A new approach of designing Multi-Agent Systems

With a practical sample

Sara Maalal
Team of Systems' Architecture, Laboratory of computing, Systems and Renewable Energy
National and High School of Electricity and Mechanic
ENSEM BP 8118, Oasis
Casablanca, Maroc

Malika Addou
Team of Systems' Architecture Laboratory of computing, Systems and Renewable Energy
Hassania School of Public
Works EHTP BP 8108, Oasis
Casablanca, Maroc

***Abstract*—Agent technology is a software paradigm that permits to implement large and complex distributed applications [1]. In order to assist analyzing, conception and development or implementation phases of multi-agent systems, we've tried to present a practical application of a generic and scalable method of a MAS with a component-oriented architecture and agent-based approach that allows MDA to generate source code from a given model. We've designed on AUML the class diagrams as a class meta-model of different agents of a MAS. Then we generated the source code of the models developed using an open source tool called AndroMDA. This agent-based and evolutive approach enhances the modularity and genericity developments and promotes their reusability in future developments. This property distinguishes our design methodology of existing methodologies in that it is constrained by any particular agent-based model while providing a library of generic models [2].**

Keyword- Software agents; Multi-agents Systems (MAS); Analysis; Software design; Modeling; Models; Diagrams; Architecture; Model Driven Architecture (MDA); Agent Unified Modeling Language (AUML); Agent Modeling Language (AML).

I. INTRODUCTION

Currently the computer systems are increasingly complex, often distributed over several sites and consist of software interacting with each other or with humans. The need for model human behavior in specific computer programs has prompted officials to use technology that affected the last decade and whose movements are very remarkable. In this context, designing multi-agent systems (MAS) is complex because they require the inclusion of several parts of the system which can often be approached from different angles. We must identify and analyze all system problems to find models for multi-agents to implement and integrate them into a coherent system. This is the software engineering and well justifies the use of a method of analysis, design and development of multi-agents systems [2].

This paper describes a practical example of a new generic model designed for modeling multi-agent systems and based on a class diagram, defining the different types of agents and meeting our needs for development and testing of MAS applications.

II. MULTI-AGENT SYSTEMS

A. Definitions

- An agent is a computer system within an environment and with an autonomous behavior made for achieving the objectives that were set during its design [3].

- A multi-agents system is a system that contains a set of agents that interact with communications protocols and are able to act on their environment. Different agents have different spheres of influence, in the sense that they have control (or at least can influence) on different parts of the environment. These spheres of influence may overlap in some cases; the fact that they coincide may cause dependencies reports between agents [4].

The MAS can be used in several application areas such as e-commerce, economic systems, distributed information systems, organizations...

B. Types of agent

Starting from the definitions cited above, we can identify the following agent types [5]:

- The reactive agent is often described as not being "clever" by itself. It is a very simple component that perceives the environment and is able to act on it. Its capacity meets mode only stimulus-action that can be considered a form of communication.
- The cognitive agent is an agent more or less intelligent, mainly characterized by a symbolic representation of knowledge and mental concepts. It has a partial representation of the environment, explicit goals, it is capable of planning their behavior, remember his past actions, communicate by sending messages, negotiate, etc..
- The intentional agent or BDI (Belief, Desire and Intention) is an intelligent agent that applies the model of human intelligence and human perspective on the world using mental concepts such as knowledge, beliefs, intentions, desires, choices, commitments. Its behavior can be provided by the award of beliefs, desires and intentions.

- The rational agent is an agent that acts in a manner allowing it to get the most success in achieving the tasks they were assigned. To this end, we must have measure of performance, if possible objective associated with a particular task that the agent should run.
- The adaptive agent is an agent that adapts to any changes that the environment can have. He is very intelligent as he is able to change its objectives and its knowledge base when they change.
- The communicative agent is an agent that is used to communicate information to all around him. This information can be made of his own perceptions as it may be transmitted by other agents.

Figure 1. Types of agents

III. The Design Methodologies – State of the art

Building high quality software for real-world applications is a difficult task because of the large number and the flexibility of components but also because of the complexity of interconnections required. The role of software engineering is precisely that of providing methodologies that can facilitate control of this complexity. A methodology by definition can facilitate the process of engineering systems. It consists of guides that cover the entire lifecycle of software development. Some are technical guides; others are managing the project [6].

We'll name "method" the approach to use a rigorous process for generating a set of models that describe various aspects of software being developed using a well- defined notation.

To this end, several software engineering paradigms have been proposed, such as object-oriented design patterns, various software architectures. These paradigms fail especially when it concerns the development of complex distributed systems for two reasons: the interactions between the various entities are defined in a too rigid way and there is no mechanism complex enough to represent the organizational structure system [7]. The paradigm of agents and multi-agent systems can be a good answer to these problems, because the agent-oriented approaches significantly increase our ability to model, design and build complex distributed systems [8].

There are many methodologies for analysis and design of multi-agent systems. We cite below some examples of existing methodologies [2]:

- The AAII methodology was developed based on the experience accumulated during the construction of BDI systems. In this methodology, we have a set of templates that, when they have been fully elaborated, define the specifications of agents such as desires, beliefs and intentions [9].
- The first version of Gaia methodology, which modeled agents from the object-oriented point of view, was revisited 3 years later by the same authors in order to represent a MAS as an organized *society* of individuals [10]. In fact, the agent entity, which is a central element of the meta-model of Gaia, can play one or more roles. A role is a specific behavior to be played by an agent (or kind of agents), defined in term of permissions, responsibilities, activities, and interactions with other roles. When playing a role, an agent updates its behavior in terms of services that can be activated according to some specific pre- and post- conditions. In addition, a role is decomposed in several protocols when agents need to communicate some data. The environment abstraction specifies all the entities and resources a multi-agent system may interact with, restricting the interactions by means of the permitted actions [1].

  The Gaia methodology gives the possibility to design MAS using an organizational paradigm and to traverse systematically the path that begins by setting out the demands of the problem and to lead to a fairly detailed and immediate implementation [9]. Gaia permits to design a hierarchical non-overlapping structure of agents with a limited depth. From the organizational point of view, agents form teams as they belong to a unique organization, they can explicitly communicate with other agents within the same organization by means of collaborations, and organizations can communicate between them by means of interactions. If inter-organization communication is omitted, coalitions and congregations may also be modeled [1].

  However, this methodology is somewhat limited since we can describe MAS with different architectures of agents [9].
- The main contribution of MESSAGE was the definition of meta-models for specification of the elements that can be used to describe each of the aspects that constitute a multi-agent system (MAS) from five viewpoints: organization, agents, goals/tasks, interactions and domain. MESSAGE adopted the Unified Process and centered on analysis and design phases of development [11].
- INGENIAS starts from the results of MESSAGE and provides a notation to guide the development process of a MAS from analysis to implementation [12] [13].

  It is both a methodology and a set of tools for development of multi-agent systems (MAS). As a methodology, it tries to integrate results from other proposals and considers the MAS from five complementary viewpoints: organization, agent,

tasks/goals, interactions, and environment. It is supported by a set of tools for modeling (graphical editor), documentation and code generation (for different agent platforms). The INGENIAS methodology does not explicitly model social norms, although they are implicit in the organizational viewpoint. Organizational dynamics are not considered i.e., how agents can join or leave the system, how they can form groups dynamically, what their life-cycle is, etc [14]. The authors have developed an agent-oriented software tool called INGENIAS Development Kit (IDK) [15]. It allows to edit consistent models (according to INGENIAS specification) and to generate documented code in different languages such as JADE [16], Robocode, Servlets or Gracias Agents [1].

- Multi-agent systems Software Engineering (MaSE) is a start-to-end methodology that covers from the analysis to the implementation of a MAS [17]. The main goal of MaSE is to guide a designer through the software life-cycle from a documented specification to an implemented agent system, with no dependency of a particular MAS architecture, agent architecture, programming language, or message-passing system.
- AUML (Agent Unified Modeling Language) is an evolving standard for a design methodology to support MAS. It is based on the UML methodology used with object oriented systems. This notation was proposed to adapt the UML's one in order to describe the agent-oriented modeling [18].

AUML provides tools for:

- Specification protocol of interaction between agents,
- Representation of the internal behaviour of an agent,
- Specification of roles, package interface agent, mobility, etc [2].

- The Agent Modeling Language (AML) is a semiformal visual modeling language for specifying, modeling and documenting systems that incorporate concepts drawn from multi-agents systems (MAS) theory [19].
- ASPECS (Agent-oriented Software Process for Engineering Complex Systems) provides a holonic perspective to design MAS [20]. Considering that complex systems typically exhibit a hierarchical configuration, on the contrary to other methodologies, it uses *holons* instead of atomic entities. Holons, which are agents recursively composed by other agents, permit to design systems with different granularities until the requested tasks are manageable by individual entities.

  The goal of the proposed meta-model of ASPECS is to gather the advantages of organizational approaches as well as of those of the holonic vision in the modeling of complex system [1].

All these methodologies presented above are still quite recent. They are mainly focused on the analysis phase, whereas design and implementation phases are missing or are redirected to agent-oriented methodologies, which do not offer enough tools to model organizational concepts. Therefore, there is still a gap between analysis and design, which must be specified clearly, correctly and completely [14].

Finally, the maturity of methodologies can be analyzed by the number of systems that have adopted them. Most of analyzed methodologies have associated applications that show their feasibility. These methodologies have been applied in different fields such as medical informatics [21], manufacturing [20] [22], and e-commerce [23]. MaSE and INGENIAS are the most used ones. Unfortunately, the number of real world applications that use agent-oriented methodologies is still low [1].

IV. The MDA Approach

The MDA (Model Driven Architecture) proposes a methodological framework and architecture for systems development that focuses first on the functionality and application behavior, without worrying about the technology with which the application will be implemented. The implementation of the application goes through the transformation of business models in specific models to a target platform (Fig.2). One research was done in this area as the dissertation of Jarraya T. [24]

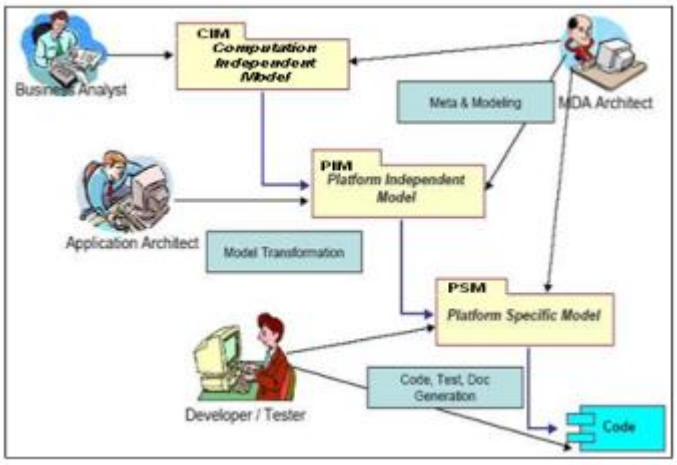


Figure 2. The MDA approach

The business process independent of automation, which comes from the expression of need, is described as a **"CIM" (Computation Independent Model).** The detailed functional analysis, the heart of the process is concentrated in the **"PIM" (Platform Independent Model)**, which, as its name suggests, is strictly independent of the technical architecture and the target language. The **"PSM" (Platform Specific Model)** is the model for engineering design obtained by transformation of PIM by projection on the target technical architecture. It is this model that is based on code generation [5].

The benefits to businesses on the MDA are primarily:

- The fact that architectures based on MDA are ready for technological developments.
- The ease of integrating applications and systems around a shared architecture
- Broader interoperability for not being tied to a platform.

One of the main tools of MDA, we have AndroMDA who takes as its input a business model specified in the Unified Modeling Language (UML) and generates significant portions of the layers needed to build, for example, a Java application [25]. AndroMDA's ability to automatically translate high-level business specifications into production quality code results in significant time savings when implementing Java applications. The diagram below maps various application layers to, for examples, Java technologies supported by AndroMDA [5].

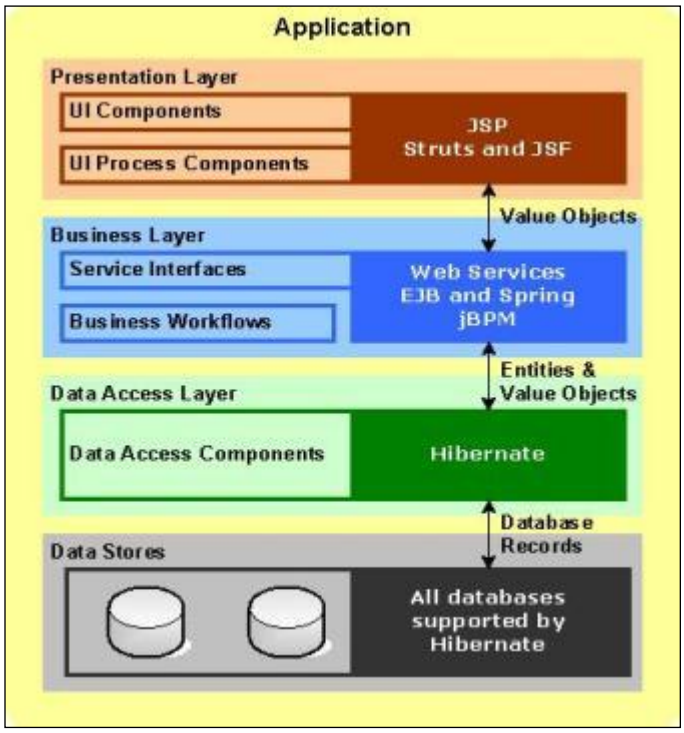


Figure 3. Application layers supported by AndroMDA

- **Presentation Layer:** AndroMDA currently offers two technology options to build web based presentation layers: Struts and JSF. It accepts UML activity diagrams as input to specify page flows and generates Web components that conform to the Struts or JSF frameworks.
- **Business Layer:** The business layer generated by AndroMDA consists primarily of services that are configured using the Spring Framework. These services are implemented manually in AndroMDA-generated blank methods, where business logic can be defined. These generated services can optionally be front-ended with EJBs, in which case the services must be deployed in an EJB container (e.g.,JBoss). Services can also be exposed as Web Services, providing a platform independent way for clients to access their functionality. AndroMDA can even generate business processes and workflows for the jBPM workflow engine (part of the JBoss product line).
- **Data Access Layer:** AndroMDA leverages the popular object-relational mapping tool called Hibernate to generate the data access layer for applications. AndroMDA does this by generating Data Access Objects (DAOs) for entities defined in the UML model. These data access objects use the Hibernate API to convert database records into objects and vice-versa. AndroMDA also supports Enterprise Java Beans EJB3/Seam [26] for data access layer (pre-release).
- **Data Stores:** Since AndroMDA generated applications use Hibernate to access the data, you can use any of the databases supported by Hibernate.

The generation process of AndroMDA is as follows [5] :

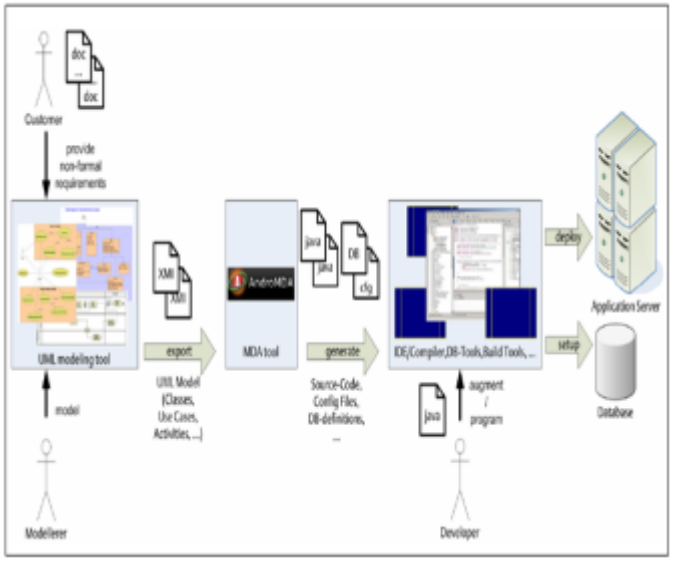


Figure 4. Generation process of AndroMDA

- Preparation of the project in MagicDraw
- Preparing use cases
- Preparation of class diagram
- Preparation of state charts
- Code Generation
- Generating the database
- Deploy the application

V. Proposed Approach

Our approach is based on model driven architecture (MDA) which aims to establish the link between the existing agent architectures and models or meta-model multi-agent systems that we build based on AUML. Our idea is to offer a design methodology based on agents AUML notation for establishing a generic class diagram that the designer can use to design his system [3]. This diagram is considered as a meta-model which

is not generated by any tool and must be defined by the modeler himself.

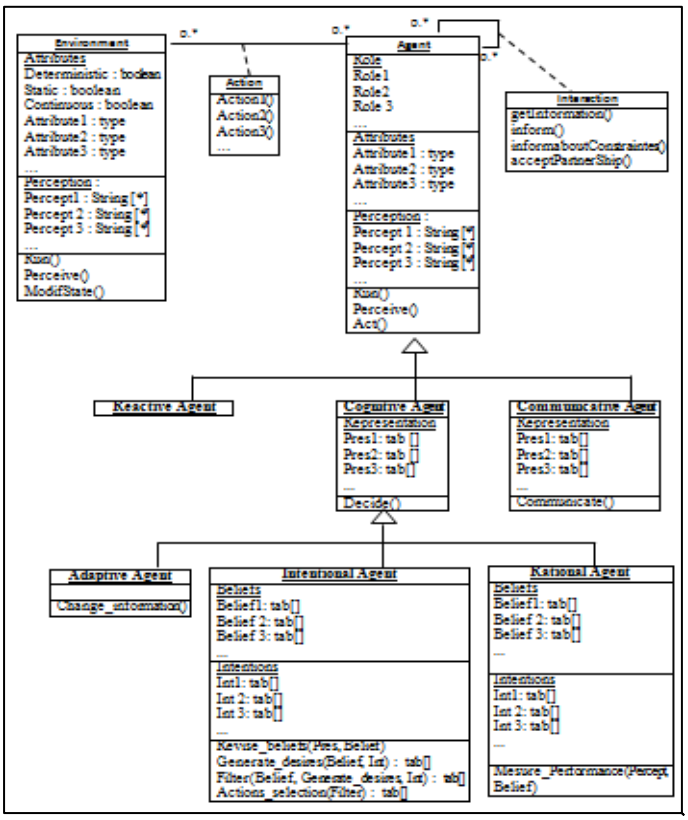


Figure 5. An AUML generic class diagram for a MAS

Our approach has a lot of benefits, it allows:

- Reducing costs and development times for new applications.
- Improving quality of applications.
- Reducing complexity of application development.
- Ability to generate all the necessary components described.
- Modularity and reusability of the developments.
- Coercion by the MDA model.
- Generating a library of generic models.

A. *Description of the AUML generic Class Diagram*

The diagram is conceived in three layers, each one is represented by a relationship between classes: A first part which is a relation between agent and its environment, a second part of specialisation of the agent class, and at the last part, a specialisation of the cognitive agent class [3].

1- The first part

The first part consists of two important classes:

- Environment,
- Agent

- **Environment** is an important class on the diagram because it influences all the system. Environment's data is represented by two sections, Attributes and Perceptions. Attributes can be all the information that an environment should have, plus the following common information:
  - *Deterministic* when the next state of the environment is determined in a unique way by the current state and action of the agent, so the environment is deterministic. If the outcome is uncertain (especially if, as a result of action of the agent, the environment can evolve in different ways), we are in the non-deterministic case.
  - *Static* if the environment cannot change its state without the intervention of the agent. The environment is dynamic if its state can change without the action of the agent in the time interval between two perceptions of the agent.
  - *Continuous* if any portion of an environment state to another requires passing through a sequence of intermediate states, otherwise the environment is discrete.

  Perception is a section where the designer should determinate all environment perceptions, example: number of agents.

  Environment contains several functions allowing to start running, to perceive information from agents linked to it and to modify its state after each action from those agents, that is respectively Run(), Perceive() and ModifState().

- **Agent** is the main class on the diagram that allows the designer to express all agent properties. The constructor of Agents takes three sections: Roles, Attributes and Perception. Roles are agent functionalities. Attributes are all information that an agent should possess. And finally Perception which is a section where the designer should determinate all agents' perceptions about his environment or the other agents.

  Agent contains several functions who allows starting running and perceiving information from environment or agents linked to it and to execute all its actions, that is respectively Run(), Perceive() and Act().

The first part consists also of two important association classes:

-Action, between agent and his environment.
-Interaction, between agents.

- **Action** is an association class between agent and environment. It lists all possible actions that an agent can execute on his environment.

- **Interaction** is a reflexive association class between agents. Agent can request information by the getInformation() function and send it by the inform() function. Agent may also deal with some constraints

that it is possible to inform by the function informaboutConstraintes(). The acceptance of partnership is added also to the main functionalities of Agent by the function acceptPartnerShip().

2- The second part

The second part represents a specialisation relation of the Agent class. It consists of three important classes:

- Reactive agent,
- Cognitive agent,
- Communicative agent.

- **Reactive agent is** a type of agent. It possesses the same properties of the Agent class.

- **Cognitive agent is** another specialization of the Agent class. In this class, the designer should determinate the representations of the agent that he must have during its execution. The class possesses also one important function "Decide()" where agent can decide to execute an action or not according to his goals.

- **Communicative agent** is the last specialization of the Agent class. Like Cognitive agent class, Communicative agent class has representations but possesses a different function called "Communicate()" where agent must use to communicate his information to the other agents.

3- The third part

The third part represents a specialization relation of the Cognitive agent class. It consists of three important classes:

- Adaptive agent,
- Intentional agent,
- Rational agent.

- **Adaptive agent** is a type of cognitive agent. It possesses the same properties of the Agent class, the knowledge base and the "Decide()" function. As mentioned in the types of agent section above, an adaptive agent is able to change its objectives and its knowledge base as and when these changes. This functionality is expressed by the "Change_information()" function.
- **Intentional agent or BDI Agent is** designed from the "Belief-Desire-Intention" model. It is a type of cognitive agent. In the same case of Adaptive Agent class, this class possesses the same properties of the Agent class, the knowledge base and the "Decide()" function.
  In this class, the designer should determinate the agent's beliefs represented by the Beliefs section. The beliefs of an agent are the information that the agent has on the environment and other agents that exist in the same environment. Beliefs may be incorrect, incomplete or uncertain, and because of that, they are different from knowledge of the agent, which is information still true. Beliefs can change over time as the agent by its ability to perceive or interact with other agents, collects more information.
  The designer should also determinate the agent's intentions represented by the Intentions section. The intentions of an agent are the actions it has decided to do to accomplish his goals.
  To choose the correct agent's beliefs from the incorrect ones, this class offers the "Revise_beliefs(Pres, Belief)" function which is based on the agent's knowledge base and his beliefs. Then, the "Generate_desires(Belief, int)" function comes to generate all the agent's desires that he may be able to accomplish at once. The desires of an agent representing all things the agent would like to see made. An agent may have conflicting desires, in which case he must choose between her desires a subset that is consistent. This subset consists of his desires is identified with the beliefs and the intentions of the agent.
  Another function comes after that, the "Filter(Belief, Generate_desires, int)" which filters all those elements above and gives the consistent beliefs, desires and intentions of the intentional agent.
  Finally, the agent can select his actions according to this filtration and execute them by the "Actions_selection(Filter)" function.

- **Rational agent** is the last specialisation of the Cognitive Agent class. Like Intentional Agent class, Rational Agent class has the Beliefs and the Intentions sections but possesses just one function called "Mesure_performance(Percept, Belief)" where agent must use to execute his actions as efficient as possible. This function is based both on his perceptions and his beliefs.

B. The generic UML Class Diagram

This generic AUML class diagram was subsequently converted into a generic class diagram based on UML notation. This transformation will allow the designer to easily use AndroMDA to generate the source code equivalent to its UML diagram [1].

The passage from AUML to UML was performed by following the steps below:

1. Keep the same titles of classes and associations which constitute the AUML diagram.
2. Assign roles, perceptions, intentions, beliefs and representations of each agent, and any possible additional attributes, in the attributes part of the UML class.
3. Combine all methods or functions in the operations part of the UML class.

We can obtain, in the end, the following result shown in Fig. 6:

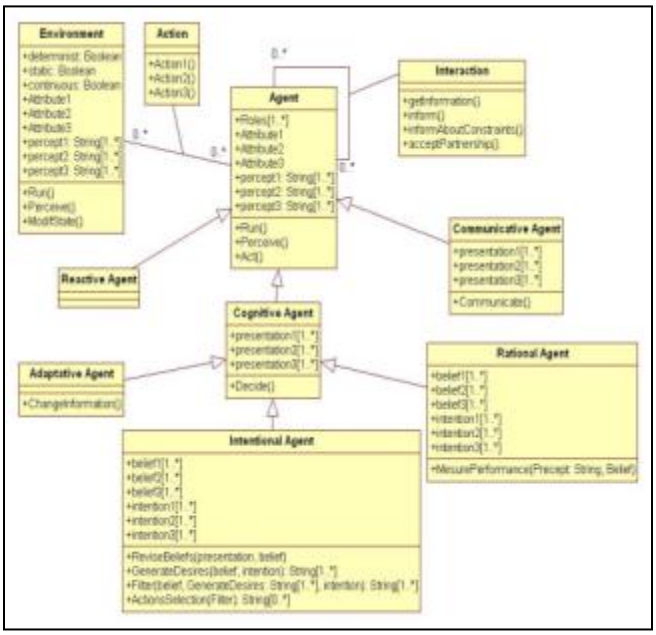


Figure 6. An UML generic class diagram for a MAS

Our approach can present one desadvantage. It is the complexity of generating a good code source by AndroMDA. The model developed at the design phase, should be reliable in order to build the application and realize its implementation without errors [5].

V. Application ExAmple

A. *Description*

Our proposed AUML class diagram was used for design of one multi-agent system for a Chat Application. This example is designed as follows [5]:

- Three reactive agents: These agents will be the chatters, the interest that these are reactive agents relies on the fact that an agent doesn't react before the declaration of the name of the receiver by the user of the application. Therefore an agent will react to get ready to catch the name and the message and to send it to the appropriate person. He will react also to clear the sent and the received message from their area in his interface.

We can respectively obtain the following AUML and UML diagrams corresponding to this example, shown in the Figures 7 and 8:

B. *Realization*

To validate our model for this example, we've tried to download AndroMDA with all the required dependencies (including all profiles referenced by models). Then, we generated our project « ChatAgents » by running « **mvn org.andromda.maven.plugins:andromdaapp-maven plugin:3.4-SNAPSHOT:generate** ». The result of this command is as follows:

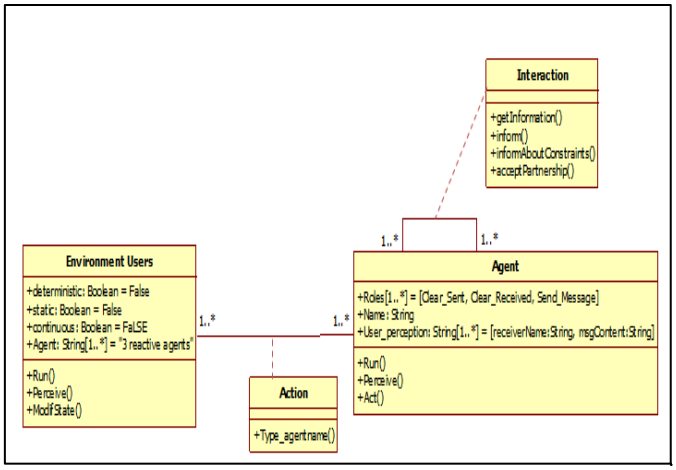


Figure 7. AUML Class diagram for a chat application

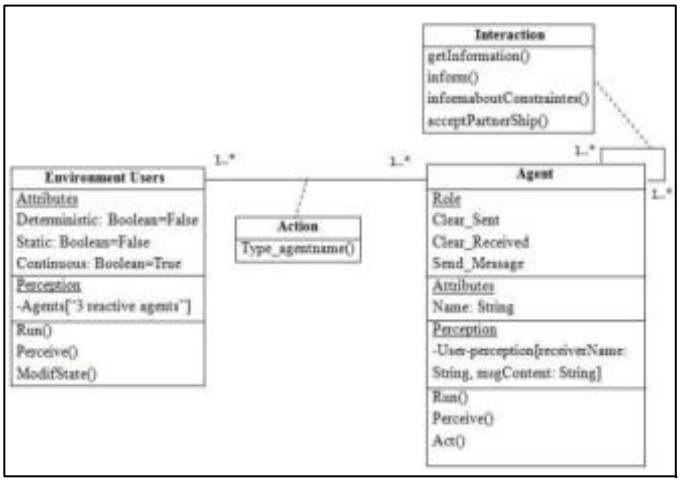


Figure 8. UML Class diagram for a chat application

When we examine the various folders and files created by the andromdapp plug-in, we will notice files called pom.xml in various folders under ChatAgents. These files make up several Maven projects. In fact, the ChatAgents directory contains a hierarchy of Maven projects as shown below [5].

- **ChatAgents:** This is the master project that controls the overall build process and common properties.
- **mda:** The mda project is the most important sub-project of the application. It houses the ChatAgents UML model under the src/main/uml directory. The mda project is also where AndroMDA is configured to generate the files needed to assemble the application.
- **common:** The common sub-project collects resources and classes that are shared among other sub-projects. These include value objects and embedded values.

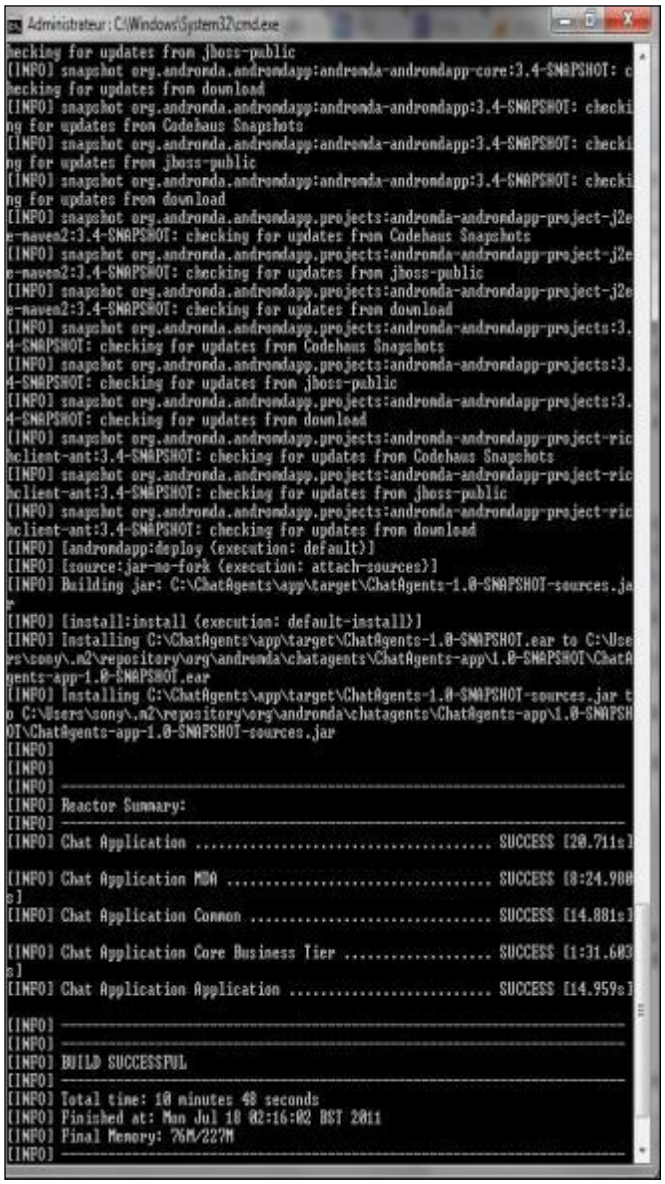


Figure 9 : ChatAgents project generation

```
ChatAgents
 |
 |-- mda
 |
 |-- common
 |
 |-- core
 |
 |-- web
 |
 +-- app
```

- **core:** The core sub-project collects resources and classes that use the Spring framework, optionally making use of Hibernate and/or EJBs under the hood. These include entity classes, data access objects, hibernate mapping files, and services.
- **web:** The web sub-project collects those resources and classes that make up the presentation layer.
- **app:** The app sub-project collects those resources and classes that are required to build the .ear bundle.

By opening the file **"ChatAgents.xml"** in MagicDraw, we will be able to build various graphs of our model to generate then the source code of the entire application. Note that AndroMDA can't read MagicDraw 17 models directly. Therefore, you can export it to another file format: EMF-UML2.

After import of AndroMDA profiles to use for our application, we designed our class diagram as shown in Fig.10 as follows [5]:

The result of exporting our **"ChatAgents"** model to EMF-UML2 format is located in the folder C:/ChatAgents/mda/src/main/uml in explorer. Below his content:

- **ChatAgents.xml:** the MagicDraw 17 model file.
- **ChatAgents.uml:** ChatAgents model in EMF/UML2 format. It's the file that will be processed by AndroMDA.
- 10 files ending with .profile.uml: the different profiles used by ChatAgents.uml

Following the definition of our model, the generation of application code is achieved by executing the command "**mvn install**", the result appears as in the figure [5].

Thus, the class **"Chat.java"** is created and can be easily accessed and modified by the developer where he has the ability to implement its operations in the generated code.

We conducted this implementation and got the final result.

VI. Conclusion And Future scope

The purpose of this paper is to demonstrate the feasibility of our approach to analyze, design and implement multi-agent systems. With AUML modeling and MDA, we can generate all the necessary components described by the class meta-model that we proposed. Which leads us to obtain a generic design based on SOA more or less reusable components using one of the most MDA tools used in development is AndroMDA [27].

In the future, we would like to model another application sample of our model but in a more complex form using cognitive or adaptive agents and in other platforms like C++, Web services, etc. It will help us to validate the efficacy of our proposed approach and lead us to consider it as a generic approach which can be adopted by every type of information system and used for any real world application.

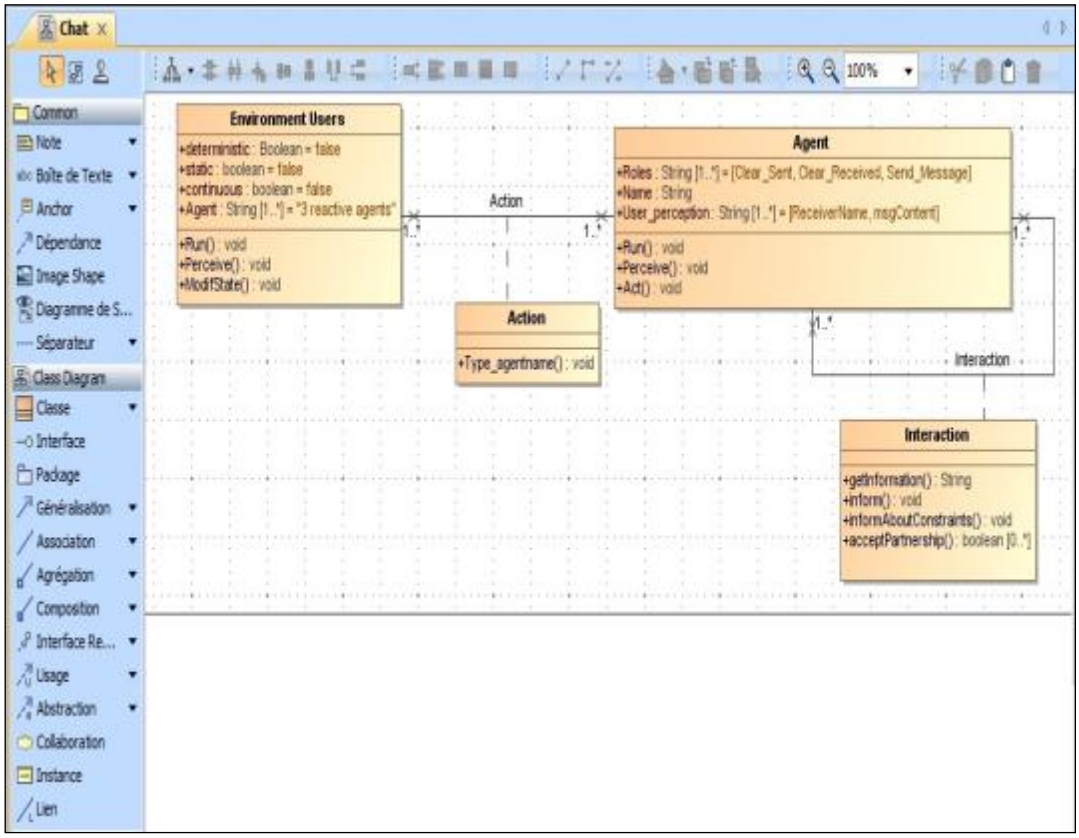


Figure 10 : Class diagram built on MagicDraw 17

Figure 11 : Code generation after definition model

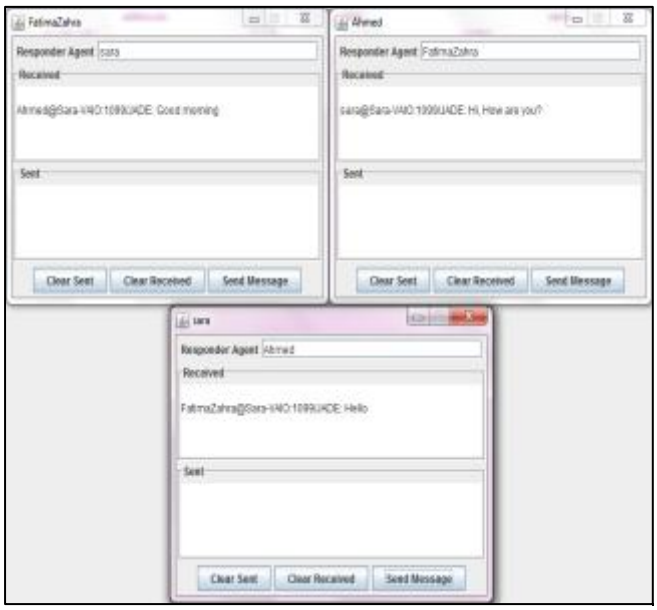


Figure 12. Chat application with three agents

ACKNOWLEDGMENT

I would like to thank to my advisor Ms. M. Addou, Phd. for his invaluable guidance and many useful suggestions during my work on this paper. I would also like to express my gratitude to all those who gave me the possibility to complete this paper.

REFERENCES

[1] D. Isern, D.Sanchez, A.Moreno, "Organizational structures supported by agent-oriented methodologies", The journal of Systems and Software, vol. 84, n. 2, Oxford, UK: Elsevier, 2011, pp. 169-184.

[2] S. Maalal, M. Addou, "A Model Design of Multi-Agent Systems", Proceedings of the 2nd Edition of the IEEE International Conference on Multimedia Computing and Systems ICMCS'11, Ouarzazate Morocco, p. 674, 2011.

[3] M. Wooldridge, Intelligent Agents, Multi agent systems, In The MIT Press, "A modern Approach to Distributed Artificial Intelligence", (England Massachutts London: MIT Press Cambridge, 1995, p. 27-78)

[4] M. Wooldridge, An Introduction to Multi-Agent Systems, Wiley & Sons, 2000.

[5] S. Maalal, M. Addou, "A practical application of a method of designing multi-agent systems based on the AUML language and the MDA approach", Proceedings of the Fourth Workshop on Information Technologies and Communication WOTIC'11, Casablanca, Morocco, p.104, 2011.

[6] O. Shehory, A. Sturm, "Evaluation of modeling techniques for agent-bases systems", Proceedings of the 5th International Conference on Autonomous Agents, pp.624-631, 2001.

[7] N. R. Jennings, "On agent-based software engineering", Artificial Intelligence, vol. 117, pp. 277-296, 2000.

[8] M. Wooldridge, N. R. Jennings, "Intelligent agent: Theory and practice", The Knowledge Engineering Review, Vol. 10, n. 2, pp. 115-152, 1995.

[9] A. M. Florea, D. Kayser, S. Pentiuc, A. El Fallah Segrounichi, Intelligents agents, Agents Intelligents, Politechnica University of Bucharest, 2002.

[10] L. Cernuzzi, T. Juan, L.Sterling, F. Zambonelli, "The Gaia methodology: basic concepts and extensions", Methodologies and Software Engeneering for Agent Systems, US: Springer, pp.69-88, 2004.

[11] J. Pavón, , J. Gómez-Sanz., "Agent Oriented Software Engineering with INGENIAS", Proceedings of the international Central and Eastern European conference on Multi-Agent Systems CEEMAS'03, pp.394-403, 2003.

[12] R. Fuentes-Fernández, I. García-Magariñio, A.M. Gómez-Rodríguez, J.C. González-Moreno, "A technique for defining agent-oriented engineering processes with tool support", Artificial Intelligence, vol.23, pp.432-444.

[13] J. Pavón, , J.J. Gómez-Sanz., R. Fuentes, 'The INGENIAS methodology and tools" in Agent-oriented Methodologies, B. Henderson-Sellers and P. Giorgini Eds. Idea Group, 2005, pp. 236–276.

[14] E. Argente, V. Julian, V. Botti, "Multi-agent system development based on organizations", Electronic Notes in Theoretical Computer Science, vol.150, pp.55-71, 2006.

[15] IDK (INGENIAS Development Kit), http://sourceforge.net/projects/ingenias/

[16] JADE (Java Agent DEvelopment Framework), http://jade.tilab.com/.

[17] S.A. DeLoach, "The MaSE methodology", in Methodologies and Software Engineering for Agent Systems, F. Bergenti, M.P Gleizes, F. Zambonelli, Eds. The Agent-oriented Software Engineering Handbook. Kluwer Academic Publishers, 2004, pp. 107–125

[18] S. Lynch, K. Rajendran, "Design Diagrams for Multi-agents Systems", Proceedings of the 16th Annual Workshop of the Psychology of Programming Interest Group PPIG'04, pp. 66-78, 2004.

[19] R. Cervenka, I. Trencansky, "Agent Modeling Language (AML): A Comprehensive Approach to Modeling MAS", Informatica, vol. 29, n. 4, pp. 391-400, 2005.

[20] M. Cossentino, N. Gaud, V. Hilaire, S.Galland, A. Koukam, 'ASPECS: An Agent-oriented Software Process for Engineering Complex Systems: How to design agent societies under a holonic perspective", 2010.

[21] D. Isern, C. Gómez-Alonso, A. Moreno, "Methodological development of a multi-agent system in the healthcare domain", Commun, SIWN 3, pp. 65–68, 2008.

[22] A. Giret, V. Botti, S. Valero, "MAS methodology for HMS", In the Second International Conference on Industrial Applications of Holonic and Multi-Agent Systems HoloMAS, Springer-Verlag, Copenhagen, Denmark, pp. 39–49, 2005.

[23] J. Ferber, O. Gutknecht, F. Michel, "From agents to organizations: an organizational view of multi-agent systems", in Springer-Verlag Berlin Heidelberg, P. Giorgini, J. Müller, J. Odell, Eds 2003, in the 4th International Workshop on Agent-oriented Software Engineering IV (AOSE), Melbourne, Australia, pp. 214–230, 2003.

[24] T. Jarraya, Re-use of interaction protocols and Career-oriented models for multi-agents development, Réutilisation des protocoles d'interaction et Démarche orientée modèles pour le développement multi-agents , Ph.D. Thesis, Dept. Computer Engineering, University of Reims Champagne Ardenne, France, 2006.

[25] N. Bhatia, "Getting Started with AndroMDA for Java" (www.andromda.org, 2010).

[26] JBoss Seam (http://www.jboss.com/products/seam/).

[27] S. Maalal, M. Addou, "A Model Design of Multi-Agents Systems", in the International Conference on Models of Information and Communication Systems MICS'10, Rabat, Morocco, 2010, unpublished.

AUTHORS PROFILE

Sara Maalal was born in Rabat the Morocco's capital in 1985. She received his professional master in Computer Engineering and Internet (3I), Option: Security Networks and Systems, in 2008 from the Faculty of science of HASSAN II University, Casablanca, Morocco. In 2010 she joined the system architecture team of the National and High School of Electricity and Mechanic (ENSEM: Ecole Nationale Supérieure d'Electricité et de Mécanique), Casablanca, Morocco.

Her actual main research interests concern Designing and modeling Multi-Agent Systems.

Ms. Maalal is actually a Software Engineer in a Moroccan multinational society called Hightech Payment Systems (HPS) which has always proved itself as a leading payment solutions provider.

Malika Addou received her Ph.D. in Artificial Intelligence from University of Liege, Liege, Belgium, in 1992. She got her engineer degree in Computer Systems from the Mohammadia School of Engineers (EMI : Ecole Mohammadia des ingénieurs), Rabat, Morocco in 1982. She is Professor of Computer Science at the Hassania School of Public Works (EHTP : Ecole Hassania des Travaux Publics), Casablanca, since 1982.

Her research focuses on Software Engineering (methods and technologies for design and development), on Information Systems (Distributed Systems) and on Artificial Intelligence (especially Multi-Agent Systems technologies).