Current Awareness Bulletin

of

SCHOLARLY ARTICLES PUBLISHED

BY

Faculty, Students and Alumni

~ March 2012 ~

DELHI TECHNOLOGICAL UNIVERSITY CENTRAL LIBRARY
(formerly Delhi College of Engineering, Bawana Road, DELHI)

PREFACE

This is the first Current Awareness Bulletin Service started by Delhi Technological University Library. The aim of the bulletin is to compile, preserve and disseminate information published by the Faculty, Students and Alumni for mutual benefits. The bulletin also aims to propagate the intellectual contribution of DTU as a whole to the academia. It contains information resources available in the internet in the form of articles, reports, presentation published in international journals, websites, etc. by the faculty and students of Delhi Technological University in the field of science and technology. The publication of Faculty and Students which are not covered in this bulletin may be because of the reason that either the full text was not accessible or could not be searched by the search engine used by the library for this purpose. To make the bulletin more comprehensive, the learned faculty and Students may provide their uncovered publication to the library either through email or in CD, etc.

This issue contains the information published during March 2012. The arrangement of the contents is alphabetical wise starting from A-Z. The Full text of the article which is either subscribed by the University or available in the web has been provided in this Bulletin.

CONTENTS

*Faculty
@Students/Research Scholars
# Alumni

A New Palm Print Based Fuzzy Vault System for Securing Cryptographic Key

Om Prakash Verma and Devesh Bharathan

Abstract—**Biometric unique authentication approach on encrypted message provides a totally secured information delivery without any need to hide any password or secret key. We present the usage of biometrics technique in asymmetric cryptographic domain where we protect secret key involved in cryptographic with the help of fuzzy vault created by randomizing the palm print feature with the secret key. This paper therefore uses the polynomial construction on the secret key using appropriate mapping function and taking the projection of palm print feature on the polynomial constructed which together makes the fuzzy vault. Vault can be unlocked with authentication and polynomial is reconstructed using Lagrange's interpolation. Secret key is retrieved from reconstructed polynomial and corrected with Reed & Solomon (RS) codes. The results shows that 309-bit RSA keys can be secured with palm print based fuzzy vault using the proposed system.**

Index Terms— **RSA; Lagrange's interpolation; RS codes; Galois field; PCA technique; Eigen values.**

I. Introduction

Today information security is very sensitive area of research. Cryptography is one of the most effective ways to enhance the security of the information system via its encryption and decryption modules. A secure encryption key can be associated with a biometric signature to ensure the integrity and confidentiality of communication in distributed systems. Many of the limitations of the password and PIN based encryption schemes can be alleviated by using biometric features, which are unique and can be conveniently extracted from every user. The biometric-based encryption requires physical presence of persons to be authenticated and is therefore reliable, convenient and efficient [1]. The motivation to protect secret key involved in cryptographic modules using biometric based fuzzy vault is came from the idea that current cryptographic algorithms e.g., Advanced Encryption Standard (AES) [2], Data Encryption Standard (DES) and RSA [3] have a very high proven security but they suffer from the key management problem as all these algorithms fully depend on the assumption that the keys will be kept in absolute secrecy. If the secret key is compromised, the security provided by them immediately falls apart. Another limitation of the these algorithms is that they require the keys to be very long and random for higher security, e.g., 128 bits for AES, which makes it impossible for users to memorize the keys. As a result, the cryptographic keys are stored securely (e.g., in a computer or on a smart card) and released based on some alternative authentication mechanism. If this authentication succeeds, keys can be used in encryption/decryption procedures. The most popular authentication mechanism used for key release is based on passwords. Thus, the plain text protected by a cryptographic algorithm is only as secure as the password (weakest link) that releases the correct decrypting keys. Simple passwords compromise security, but complex passwords are difficult to remember and expensive to maintain. Further, passwords are unable to provide non repudiation; a subject may deny releasing the key using password authentication.

The most remarkable work in this area is to provide cryptography based security at different stages of biometric authentication via fuzzy vault scheme [4]. This vault is a form of error-tolerant cryptographic algorithm and proved very useful in many circumstances, such as fuzzy human factor based authentication systems, where exactness of the unlock key is usually unavailable. It is inherently more reliable than password-based authentication as biometric characteristics cannot be lost or forgotten. Further, biometric characteristics are difficult to copy, share and distribute and require the person being authenticated to be present at the time and point of authentication. Hence, biometrics-based authentication is a potential candidate to replace password-based authentication, either for providing complete authentication mechanism or for securing the traditional cryptographic keys claiming that his/her password was stolen. Many of these limitations of password-based key release can be eliminated by incorporating biometric authentication.

II. Prior Work

Fuzzy vault for information security is not very new concept. It has been used for the number of years by various scientists in research work. Jules and Sudan [5] have proposed the generation of a secure vault using an un-ordered set, to lock any secret inside and referred it as fuzzy vault. The concept of fuzzy vault has been further explored by Uludag *et al.* [1] where they used fingerprint templates as an unordered set to create the vault around the secret. They further utilizes error correcting codes, such as Reed and Solomon code to produce some error tolerance in the input biometric templates, while decrypting the module. Their contribution is to hide any secret in fuzzy vault using polynomial construction under un-ordered set. The secret can

Manuscript received January 25 2012; revised March 8, 2012. This work was supported in part by the Biometric Lab, Indian Institute of Technology Delhi, Summer Internship Program.

Om Prakash Verma is with the Department of Information Technology, Delhi College of Engineering, Delhi, India (e-mail: opverma.dce@gmail.com).

Devesh Bharathan is with the Department of Information Technology, Delhi College of Engineering, Delhi, India (e-mail: devesh.bharathan@gmail.com).

be retrieved back by polynomial reconstruction, if certain points of the unordered set are known at receiving end. The security of the scheme mainly depends upon polynomial construction and reconstruction problem. They have combined the concept of fuzzy vault with biometrics (fingerprint) by using biometric template as an un-ordered set. Uludag and Jain [4] proposed the minutiae based features from the fingerprints for locking and unlocking the vault. They attempted to secure the secret key of any cryptosystem using fuzzy vault. Vault is created by taking projections on fingerprint minutiae features and Cyclic Redundancy Codes (CRC) technique is used for error correction. However, this approach is limited to its usage due to its inability to eliminate the inherent variability in minutiae feature. Nanda kumar *et al.* [6] have attempted to eliminate such variability using helper data and illustrated promising results. Feng Hao *et al.* [7] use iris biometric for generating cryptographic keys and a combination of Reed and Solomon error correcting theories for error tolerance. Clancy et *al*. [8] proposed a smart card based fuzzy vault that employed fingerprints for locking and unlocking. The presumption that acquired fingerprint images are pre-aligned is not realistic and could be the possible reason for high false rejection rate (up to 30.0%) reported in the paper. Lin and Lia [9] have done remarkable work in order to prevent repudiation but their work still required smart card and password for better implementation and hence reduces its usability. Recently, a modified fuzzy vault scheme is proposed by Feng Hao *et al*. [7] using asymmetric cryptosystem. Having generated RSA public and private keys, authors have used Reed and Solomon coding to convert the keys in to codes. Further they used two grids, one for codes and other for biometric features. The elements in the corresponding grids are in same positions. The unlocking of vault only requires the knowledge of the correct positions of the numbers in any of the grids.

Our scheme is inspired from work Fuzzy Vault for Fingerprints in [1] with a number of modifications. The fingerprint features are very difficult to extract from the elderly, laborer, and handicapped users. In compare, we explored the usage of palm print biometric to create fuzzy vault. Having a large surface area palm images are less affected with skin attributes like, amount of grease and dust on the hand of laborers, elderly people and handicapped users as lack of palm is less likely than thumb. A mapping function is suggested to map the secret data (Secret Key of RSA system) to the coefficients of polynomial. PCA technique is used to get most randomized palm feature and projections are taken on this to create the vault. During unlocking, polynomial is reconstructed with Lagrange's interpolation and polynomial coefficients are finally mapped to secret data. In our current encoding implementation, the data is encrypted by RSA cryptosystem, which is represented in Galois Field. And for error correction we have followed RS coding (Reed and Solomon) on these field elements. Firstly, the secret key is mapped to decimal values to form polynomial coefficients. Thereafter the polynomial generated is used for taking projection on the elements of images obtained by PCA technique. Thus the elements and its projection together make the vault.

Decoding implementation is done by comparing elements of another image of user with the vault generated for that user. With the matched values the reverse of encoding methodology is followed with reconstruction of polynomial being done by Lagrange's polynomial reconstruction. And hence the secret key is generated.

III. Proposed Fuzzy Vault System

Fuzzy vault scheme is a simple and novel cryptographic construction. Suppose we have a secret, which we want to share with some specific persons, but do not want to post it indiscriminately, such as a public website. One approach is to compile a set of elements *A* with the secret and publish it in an encrypted form. To extract the secret information, one needs to have an unlocked set *B,* which is close to *A,* to unlock the vault. We have suggested an alternative way of creating a Fuzzy Vault around the secret key of asymmetric cryptosystem. Secret key is mapped to decimal values (with mapping function) to form polynomial coefficients as follows;

$$K = SECRET\ KEY$$
$$X = MAP\ [K]$$

where X is the formed polynomial. Most randomized Palm features data is obtained by PCA technique as follows;

$$P = PALM\ FEATURE$$
$$Q = PCA\ [P]$$

where Q contains most randomized palm features. Projections of Q are taken over X and vault is made up of Q, R and some chaff points (not lying on polynomial X) given as

$$R = EVALUTE[X, Q]$$
$$VAULT = [Q, R] + CHAFF\ POINT$$

For unlocking, the user features are randomized and vault is unlocked with matching them with Q of vault. Polynomial is reconstructed and its coefficients are mapped back to secret key as

$$G = USER\ PALM$$
$$H = PCA\ [G]$$
$$[I, J] = UNLOCK\ [H]$$
$$L = INTERPOLATE\ [I, J]$$
$$SECRET\ KEY = M = INVERSEMAP\ [L]$$

A. Mapping Function

Any generalized injective (one to one) function can be used for mapping since its inverse give unique value. So a better option is to use logarithmic function. Therefore for each 15 characters of secret key, take decimal equivalent of these values and then take log at base 2 to get constant value. Repeat it for entire length of the key and get the polynomial coefficients as follow;

$$K = SECRET\ KEY$$
$$L = GROUP\ [K] \quad (FORM\ THE\ GROUP\ OF\ EACH\ 15\ CHRACTERS\ TO\ DECIMAL\ EQUIVALENT)$$
$$M = LOG\ [L] \quad (TAKE\ THE\ LOG\ AT\ BASE\ 2\ OF\ EACH\ VALUE)$$

where *M* is the constructed polynomial.

B. PCA Mapping

Fuzzy Vault consist the projection of most randomized Palm-Features over the polynomial. The Palm print recognition is done by principal component analysis (PCA) technique to generate most randomized Palm-Feature [10]. Constructed polynomial is evaluated on randomize palm feature which is obtain as follows

P = PALMFEATURE MATRIX (SIZE = 720 X 300)
(DATABASE OF 100 PERSONS)
B = P X P' (SIZE = 720 X 720)
C = EIG (B) (SIZE = 720 X 100) (TAKE 100 EIGEN VECTORS OF B)
Q = C' X P (SIZE = 100 X 300)
P' & *C'* are transpose vectors of *P* & *C* respectively.

where *Q* is the randomized palm feature.

IV. Proposed Algorithm

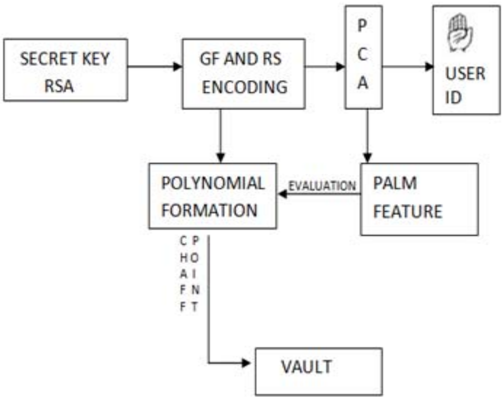


Fig. 1. Block diagram of encoding phase.

A. Locking the Vault (Encoding Phase)

Let *S* be the secret key which is generated using RSA algorithm on the given message. Suppose that *S* be the secret key of 309 characters (i.e. for strength of 512 characters), which is converted in the array format *R*. For the simplicity of our further computation, *R* is padded with some extra redundant number, so that it is of 324 lengths. Now *R* is converted to a two dimensional array *R1* of size "18x18"(square matrix). In the current encoding implementation, the array elements are represented as Galois Field [11], in order to insure that elements are taken in quantized range, so for error correction Reed and Solomon codes are applied. The Galois Field of size (2^5) and the key size is set to 18 (i.e. n = 22, k = 18, m = 5). Now RS coding is performed on Galois Field element to convert it into a two dimensional array *R2* having size (18 x 22). *R2* is converted to a one dimensional double array *A* having 396 (18 x 22) elements and also some redundant elements are padded to *A*. For polynomial formation, considering it to be of degree 26, elements of *A* is divided into groups of 15. As said earlier the mapping logarithmic function is used, therefore $\log_2$ of the decimal value of each of the 27 groups is computed and stored into an array *P1*. These groups are the coefficients of polynomial of order 26.

As discussed earlier, the PCA technique is applied to convert the given data into a (100 x 300) matrix. Using the elements of single image of each user, the projections of each value of x using the polynomial is computed. Now the x values are stored in an array *X1* and the corresponding projection is stored in another array *Y1*. Some randomly generated chaff values are added to *X1* and *Y1*, assuming genuine and chaff points are uniformly distributed. Hence *X1* and *Y1* together makes the vault locked. The scheme for the encoding phase is represented in Fig. 1.

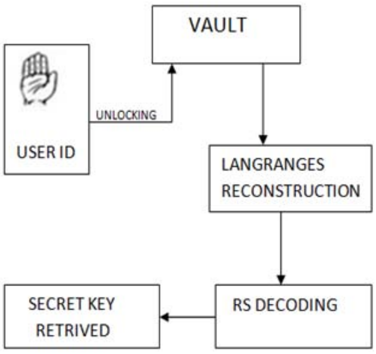


Fig. 2. Block diagram of decoding phase.

B. Unlocking the Vault (Decoding Phase):

The idea for unlocking the vault is conveniently taken by comparing the elements of second image of the user with the vault generated for that user. Now the matched values are stored in an array *W* and the corresponding projection are stored in array in an array *L*. Exactly 27 pairs is formed from *W, L* and the inverse of whole encoding procedure is followed. The Lagrange's reconstruction [12] is followed for polynomial reconstruction. Using the 27 coefficients of this polynomial, all 405 values are recomputed by first finding the anti-log of each coefficient and then getting 15 values each for a coefficient (27 x 15 = 405). These 405 values are stored in array *B1*. The inverse RS decoding is followed on the first 396 values of array *B1* taken in two dimensional array *B2* of size (18 x 22). An array *B3* of size (18 x 18) is generated with the help of RS decoding. Afterwards the secret key *S*' is retrieved from *B3* with length 309. The scheme for the encoding phase is represented in Fig. 2.

V. Experimental Results

The computation requires 37 seconds for a system with a 2.10 GHz dual processor. Further, the system is implemented in Matlab, contributing to high computational times.

The analysis of results (see Table 1) based on the Eigen values shows that false accept rate (FAR) is negligible at lower Eigen values but some significant value at higher Eigen values which can be reduce by taking more samples per user. The results obtained using the proposed method as compared to fingerprint based fuzzy vault is appreciable since we achieved 65% successful unlocking rate for a second image of user from 10 different users while Yang & Verbauwhede *et al.* [13] has given result for fingerprint with 83% successful unlocking rate for 10 prints per finger from 10 different fingers, forming a total 100 fingerprint images.

TABLE I: EXPERIMENTAL RESULTS

S:No	Test performed for Ten user's image:		
	NUMBER OF EIGEN VECTORS TAKEN	*PERCENTAGE OF VAULT UNLOCKED BY IMAGE 2 OF USER*	*PERCENTAGE OF VAULT UNLOCKED FROM IMAGE OF OTHER USER*
1.	50	20%	0%
2.	100	20%	0%
3.	300	30%	10%
4.	600	65%	20%

VI. CONCLUSION

Introduction of biometrics in cryptography domain is really a better option to make system more secure. Thus a secure encryption key can be associated with a biometric signature to ensure the integrity and confidentiality of communication in distributed systems. Many of the limitations of the password and PIN based encryption schemes can be alleviated by using biometric features, which are unique and can be conveniently extracted from every user.

We have shown that performance and security of a palm print based fuzzy vault is better option in comparable to fingerprint fuzzy vault. The experimental results from the proposed approach on the palm print images suggest its possible usage in an automated palm print based key generation system. The proposed method has inherent biometric limitations like adaptation of biometrics is not easily applicable. The proposed system is very critical in terms of precision required for proper implementation. Reconstruction of polynomial of high power with at most accuracy is somewhat very tedious job and method really lags in this phase. Therefore better polynomial formation and precision is challenge in this work.

REFERENCES

[1] U. Uludag1, S. Pankanti, and A. K. Jain1," Fuzzy Vault for Fingerprints," *Springer Berlin/Heidelberg*, vol. 3546, pp. 310-319, 2005.

[2] National Institute of Standards and Technology, *Advanced Encryption Standard(AES),* Federal Information Processing Standards Publication 197, November 26, 2001.[Online]. Available: http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf

[3] W. Stallings, *Cryptography and Network Security: Principles and Practices*, 3. Ed., Prentice Hall.,2002,ch. 1.

[4] U. Uludag and A. K. Jain, "Fuzzy fingerprint vault," in *Proc. Workshop: Biometrics: Challenges Arising from Theory to Practice*, pp. 13-16, Cambridge, UK, Aug. 2004.

[5] A. Juels and M. Sudan, "A Fuzzy Vault Scheme," in Proc. *IEEE Int'l. Symp. Information Theory*, Lausanne, pp. 408, 2002.

[6] K. Nandakumar, A. K. Jain, and S. Pankanti, "Fingerprint-based fuzzy vault: Implementation and Performance," *IEEE Trans. Info. Forensics & Security*, vol. 2, no. 4, pp. 744-757, Dec. 2007.

[7] F. Hao, R. Anderson, and J. Daugman, "Combining crypto with biometrics efficiently," *IEEE Trans. Computers*, vol. 55, pp. 1081-1088, Sep. 2006.

[8] T. C. Calancy, N. Kiyavash, and D.J . Lin, "Secure Smartcard-based Fingerprint Authentication," *ACM SIGMM , Multimedia Workshop on Biometrics Methods and Applications*, pp. 45-52, 2003.

[9] C. H. Lin and Y. Y. Lai, "A flexible biometrics remote user authentication scheme," *Computer Standards & Interfaces*, vol. 27, no. 1, pp. 19-23, Nov. 2004.

[10] A. Murat and E. Murat, "Kernel Principal Component Analysis of Gabor Features for Palmprint Recognition," *Springer Berlin / Heidelberg*, vol. 5558, 2009.

[11] Great Galois field array−MATLAB. [Online]. Availble: http://www.mathworks.in/help/toolbox/comm/ref/gf.html .

[12] H. Jeffreys and B. S. Jeffreys "Lagrange's Interpolation Formula." Methods of Mathematical Physics, 3rd ed. Cambridge, England: Cambridge University Press, p. 260, 1988.

[13] S. Yang and I. M. Verbauwhede, "Secure fuzzy vault based fingerprint verification system," *IEEE Conference, Signals, Systems and Computer*, vol. 1, pp. 577 – 581, 7-10 Nov. 2004.

[14] A. Kumar and A. Kumar," Development of a New Cryptographic Construct Using Palmprint Based Fuzzy Vault," *EURASIP Journal on Advances in Signal Processing*, 2009.

[15] W. Diffie and M. Hellman, "Multi-user cryptographic techniques," *AFIPS Proceedings* 45, pp 109–112, June 8, 1976.

[16] B. Archer, E. W. Weisstein, Lagrange interpolating polynomial, *MathWorld – A Wolfram Web Resource.*[Online]. Available: http://mathworld.wolfram.com/LagrangeInterpolatingPolynomial.html

Om Prakash Verma received his B.E. degree in Electronics and Communication Engineering from Malaviya National Institute of Technology, Jaipur, India, M.Tech. degree in Communication and Radar Engineering from Indian Institute of Technology (IIT), Delhi, India, and Ph.D. from University of Delhi, Delhi, India. From 1992 to 1998 he was assistant professor in Department of Electronics & Communication Engineering, at Malaviya National Institute of Technology, Jaipur, India. He joined Department of Electronics & Communication Engineering, Delhi College of Engineering (now Delhi Technological University) Delhi, India, as Associate Professor in 1998. Since 2007, he is head of department of Information Technology at Delhi Technological University, Delhi. He is also the author of more than 25 publications in both international journal and conference proceedings. He has guided more than 20 M.Tech. students for their theses. He has authored a book on Digital Signal Processing in 2003. He is a Principal investigator of an Information Security Education Awareness project, sponsored by Department of Information Technology, Government of India. His research interests include image processing, application of fuzzy logic in image processing, application of evolutionary algorithm in signal and image processing. (email:opverma.dce@gmail.com)

Devesh Bharathan was born in Varanasi,India, on January 11, 1989.He received B.E in Information Technology from Delhi College of Engineering (now Delhi Technological University),Delhi,India,in 2011.His reaserch interest are in Information Security and Image Processing.Presently he is working as Software Engineer in Oracle India Pvt Ltd Bangalore. (email:deveshbharathan@gmail.com)

International Journal of Pure and Applied Mathematics
Volume 75 No. 4 2012, 427-439
ISSN: 1311-8080 (printed version)
url: http://www.ijpam.eu

ijpam.eu

BOUNDARY CHARACTERISTIC ORTHOGONAL POLYNOMIALS FOR SINGULARLY PERTURBED PROBLEMS

Vivek Kumar
Department of Applied Mathematics
Delhi Technological University
Bawana Road, Delhi, 110042, INDIA

Abstract: In this paper we generate fitted mesh using the boundary characteristic orthogonal polynomials for numerically solving singularly perturbed boundary value problems. The method is based upon the orthogonal collocation method. Boundary characteristic orthogonal polynomials (BCOPs) are generated using the Gram - Schmit process from a set of linearly independent functions which also satisfy the given boundary conditions. The procedure is illustrated by taking several examples. After obtaining the fitted mesh, the problems have been solved using upwind finite difference method.

AMS Subject Classification: 65L11, 65L10, 65L50, 65M50
Key Words: boundary characteristic orthogonal polynomials (BCOPs), singularly perturbed boundary values problems (SPPs), fitted mesh methods, upwind finite difference

1. Introduction

Orthogonal polynomials have been extensively used in numerical approximations, for example, the famous Legendre, Chebyshev polynomials and many more. The importance of orthogonal projection and orthogonal decomposition,

Received: August 6, 2011

particularly in the solution of systems of linear equations and in the least square data fitting is also well known. Now a large number of books and research papers are available on orthogonal polynomials and their applications and some good references can be found in [1, 2, 3].

Problems in which a small perturbation parameter, say ϵ is multiplied to the highest derivative arise in various fields of science and engineering, for instance fluid mechanics, elasticity, hydrodynamics, etc. The main concern with such problems is the rapid growth or decay of the solution in one or more narrow "layer region(s)". These kinds of problems are known in the literature as singularly perturbed problems (SPPs).

Singular perturbation problems in consideration have shocks as boundary layers or interior layers. For such kinds of problems the solution can be smooth in most of the solution domain with small area where the solution changes very quickly. To approximate their solution it is well known (see [4, 5]) that the classical numerical methods cannot be used on uniform meshes; the reason is that the error is unbounded [6] for arbitrary values of the singularly perturbed parameter, ϵ. So when solving such problems numerically, one would like to adjust the discretization to the solution. In terms of mesh generation, we want to have many points in the area where the solution has strong variations and a few points in the area where the solution has weak variations and such method is known as fitted mesh methods.

In orthogonal collocation, we first generate the orthogonal polynomials and then we find the roots of those orthogonal polynomials (Each polynomial in an orthogonal sequence has all n of its roots real, distinct, and strictly inside the interval of orthogonality) and treat them as collocation points. The choice of collocation points is also critical and should not be arbitrary in realistic problems. The dependence of the roots according to the mesh requirement lies on the choice of the weight functions used in defining the inner product. In Section 2 we discuss how to find the BCOPs and the corresponding roots and numerical results have been presented in Section 3.

2. Generation of Mesh using BCOPs

Let us first define the inner product in the functional space for two functions $f(x)$ and $g(x)$ defined over the domain $D \in R^n$ by

$$< f, g >= \int_D w(x) f(x) g(x) dD \tag{1}$$

where $w(x)$ is the suitable chosen weight function according to the mesh defined over D. The induced norm of a function using above inner product is, therefore, given as

$$||f||^2 = \int_D w(x)f^2(x)dD. \tag{2}$$

To generate an orthogonal sequence, we can start with the set

$$\{h(x)f_i(x)\}, i = 0, 1, 2, 3.... \tag{3}$$

where $h(x)$ is the chosen function which satisfy the given boundary conditions of a differential equation and $f_i(x)$ are the linearly independent functions over the domain D. Note that each $h(x)f_i(x)$ will, therefore, also satisfy the same boundary conditions (if we have zero boundary conditions). Otherwise for non zero boundary conditions, we have to choose $f_i(x)$ which also satisfy the boundary conditions and independency.
To generate an orthogonal sequence ϕ_i, we apply the well known Gram - Schmidt process, which is given as

$$\phi_1 = hf_1 \tag{4}$$

$$\phi_i = hf_i - \sum_{j=1}^{i-1} c_{ij}\phi_j, \quad i = 2, 3, 4.... \tag{5}$$

where

$$c_{ij} =< hf_i, \phi_j > / < \phi_j, \phi_j > \tag{6}$$

The orthogonal sequence can also be normalized by dividing each ϕ_i by its norm.

2.1. BCOPs Approximations

First we try to approximate a given function, using the method of least square, which has boundary layers on both the boundaries by generating corresponding BCOPs. Suppose we take original function $F(x)$ and write it as a linear combination of generated BCOPs ϕ_is as

$$F(x) = \sum_{j=0}^{N} d_j\phi_j(x). \tag{7}$$

Then the d_js can be calculated using the concept of orthogonality (Fourier-Legendre type).

$$d_j =< F, \phi_j > / < \phi_j, \phi_j > . \tag{8}$$

Suppose we start with the function

$$F(x) = \frac{\exp((x-1)/\sqrt{\epsilon}) + \exp(-x/\sqrt{\epsilon})}{(1+\exp(-1/\sqrt{\epsilon}))} - \cos(\pi x)^2; \quad x \in (0,1), \tag{9}$$

which has boundary layers on both the sides of the interval for small values of parameter ϵ with zero boundary conditions. Let us we start with $h(x) = x(1-x)$, as it satisfies the zero boundary conditions. The inner product for this function can be defined as

$$< f, g >= \int_0^1 \frac{f(x)g(x)}{\sqrt{x(1-x)}} dx, \tag{10}$$

with weight function $w(x) = 1/\sqrt{x(1-x)}$, as it will be very helpful for generating finer mesh at the boundary points 0 and 1. As we know that the computation of the integral involved will become simpler if we deal with the polynomial functions f_is. So we start with one of the obvious choice of f_i as

$$f_i = \{1, x, x^2, x^3, ...\} \tag{11}$$

It is clear that the function hf_i also satisfies the zero boundary conditions. Using the above procedure to generate BCOPs, Some ϕ_is can be given as

$$\phi_0 = \frac{8x(1-x)\sqrt{6}}{3\sqrt{\pi}}; \tag{12}$$

$$\phi_1 = \frac{32(x^2(1-x) - (1/2)x(1-x))}{\sqrt{\pi}}; \tag{13}$$

and so on. Similarly ϕ_9 can also be given as

$$\begin{aligned}\phi_9 = &-2.330769986x(x-1)(2x-1)(3.27680 \times 10^5 x^8 - 1.310720 \times 10^6 x^7 \\ &+ 2.146304 \times 10^6 x^6 - 1.851392 \times 10^6 x^5 + 9.02144 \times 10^5 x^4 \\ &- 2.47808 \times 10^5 x^3 + 36256x^2 - 2464x + 55).\end{aligned} \tag{14}$$

For the above $F(x)$, equation (7) can be written as

$$F(x) = \sum_{j=0}^{9} d_j \phi_j(x) \tag{15}$$

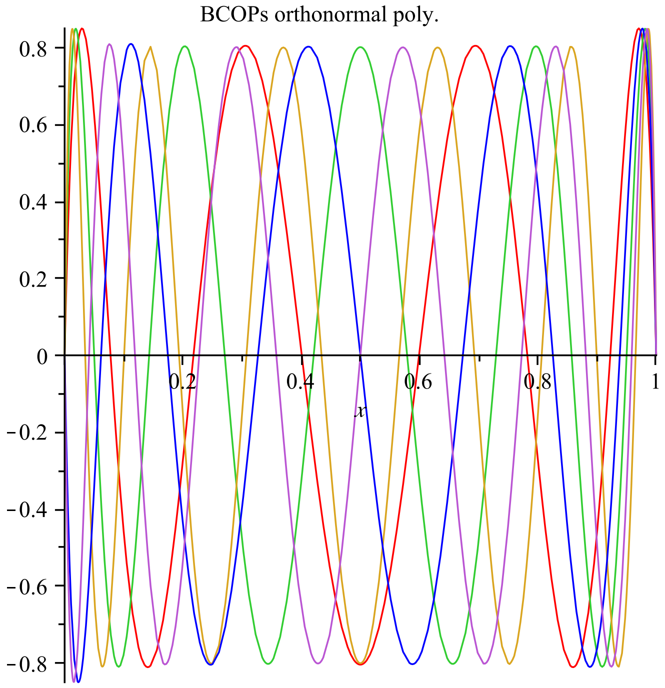


Figure 1: BCOPs for various ϕ_is

where ϕ_is as generated above. The corresponding d_is are found using orthogonal property as

$$\begin{aligned} &d_0 = -.556648, d_1 = -2.246921 \times 10^{-10}, \\ &d_2 = -.756039, ...d_{,8} = -.025290, d_9 = -1.16538 \times 10^{-13}. \end{aligned} \tag{16}$$

All the ϕ_is as used in equation (15) has been plotted in the Figure 1 and the corresponding approximated $F(x)$ is plotted in the Figure 2. As we can see in the Figure 2 that both the functions are coinciding completely with each other.

3. Numerical Results

Now we generate the fitted mesh for a given problem using the BCOPs method and solve the singularly perturbed problems using upwind finite difference methods on non-uniform meshes.

Test Problem 1. (see [7]) We consider 1D linear reaction-diffusion problem as

$$-\epsilon u''(x) + u(x) = -\cos^2(\pi x) - 2\epsilon\pi^2\cos(2\pi x), x \in (0,1), 0 < \epsilon << 1, \tag{17}$$

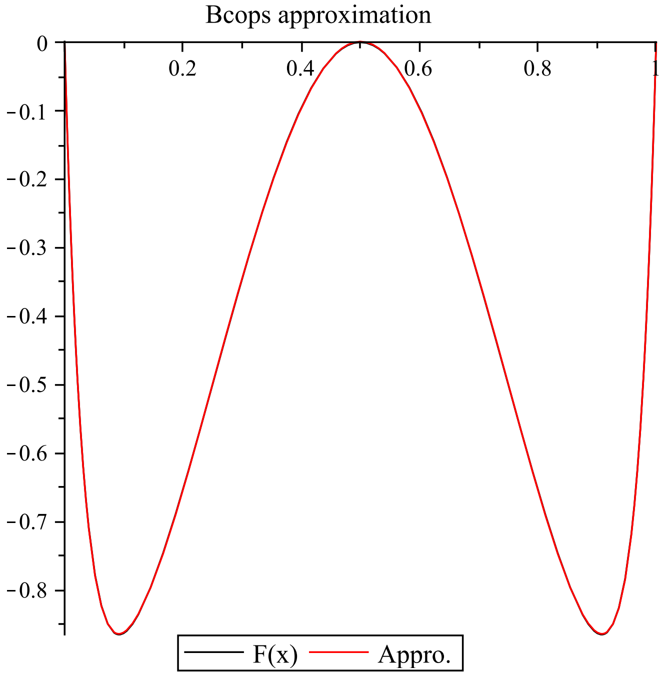


Figure 2: Approximation for $\epsilon = .001$

with boundary conditions

$$u(0) = 0,\ u(1) = 0 \tag{18}$$

whose exact solution (as also discussed in equation (9)) is

$$u(x) = \frac{\exp(-1(1-x)/\sqrt{\epsilon}) + \exp(-x/\sqrt{\epsilon})}{1 + \exp(-1/\sqrt{\epsilon})} - \cos^2(\pi x).$$

This problem has regular boundary layers of width $O(\sqrt{\epsilon})$ at $x = 0$ and $x = 1$. Using the same inner product as defined above in equation (10) we get the same orthonormal polynomials as given above equations (12 -14), we find the roots of such polynomials and apply orthogonal collocation method to compute the numerical solution. The roots of few BCOPs other than $0, 1$ are given in Table 1.

In Table 2, we have shown the maximum error for various values of perturbation parameter ϵ by using different BCOPs. Results looks better even for smaller values of ϵ. It is also clear from the Table that for each ϕ_is (say ϕ_5) maximum error is reducing as for higher degree BCOPs, the more roots are falling in the boundary layer region.

Similarly we can solve another type of problem i.e convection diffusion equation which has only one boundary layer.

$\phi_i \downarrow$	$roots \rightarrow$									
ϕ_5	.50000	.10089	.89910	.27853	.721460					
ϕ_6	.077624	.92237	.21780	.78219	.40090	.59909				
ϕ_8	.049965	.95003	.14286	.85713	.27071	.72928	.42099	.57900		
ϕ_9	.50000	.041367	.95863	.11897	.88102	.22742	.77257	.35795	.64204	
ϕ_{10}	.034806	.96519	.10054	.89945	.19348	.80651	.30731	.69268	.43428	.56571

Table 1: Roots for the corresponding ϕ_is for test problem 1

$\epsilon \downarrow \phi_i \rightarrow$	ϕ_5	ϕ_6	ϕ_7	ϕ_8	ϕ_9	ϕ_{10}
0.1	0.0706	0.0555	0.0457	0.0384	0.0330	0.0287
0.01	0.0501	0.0449	0.0392	0.0341	0.0299	0.0265
0.001	0.0288	0.0179	0.0049	0.0144	0.0244	0.0297
0.0001	0.0095	0.0156	0.0224	0.0283	0.0314	0.0308
0.00001	0.000964	0.0016	0.0026	0.0039	0.0057	0.0080

Table 2: Maximum error for various values of ϵ and mesh points for test problem 1

Test Problem 2. (see [5]) Convection diffusion equation in 1D can be given as

$$\epsilon u''(x) + 2u'(x) = 0, x \in (0,1), 0 < \epsilon << 1, \tag{19}$$

with boundary conditions

$$u(0) = 1,\ u(1) = 0 \tag{20}$$

This equation has boundary layer of order $O(\epsilon)$ at $x = 0$ and its exact solution is given as

$$u(x) = \frac{e^{-2x/\epsilon} - e^{-2/\epsilon}}{1 - e^{-2/\epsilon}}. \tag{21}$$

To find the fitted mesh (finer mesh at $x = 0$) we define the inner product as

$$< f, g >= \int_0^1 \frac{f(x)g(x)}{\sqrt{x}} dx, \tag{22}$$

with weight function as $w(x) = 1/\sqrt{x}$ to have finer mesh near $x = 0$. In this case we choose f_is as

$$f_i = \{1 - x, (1 - x)^2, (1 - x)^3, (1 - x)^4, ...\} \tag{23}$$

and $h(x) = 1 - x$. We can see that each hf_i satisfies the non-zero boundary conditions as per the requirement. Using the above procedure to generate BCOPs, we give some ϕ_is as mentioned below

$$
\begin{aligned}
\phi_0 &= \frac{\sqrt{15}(1-x)}{4};\\
\phi_1 &= -\frac{3\sqrt{5}(-7(1-x)^2+6(1-x))}{8};\\
\phi_5 &= (.26757(x-1))(7429x^5 - 14535x^4 + 9690x^3 - 2550x^2 + 225x - 3);\\
\phi_{10} &= -(0.00019037(x-1))(1.143532 10^{10} x^{10} - 5.052816 10^{10} x^9\\
&\quad + 9.427815 10^{10} x^8 - 9.669554 10^{10} x^7 + 5.9454690 10^{10} x^6 - 2.242291 10^{10} x^5\\
&\quad + 5.096116 10^9 x^4 - 6.57563 10^8 x^3 + 4.2514 10^7 x^2 - 1.049750 10^6 x + 4199);
\end{aligned}
\tag{24}
$$

and so on. Similarly we can generate much higher order BCOPs. For this example we have generated the functions upto ϕ_{14}. The roots of the corresponding BCOPs (since $x = 0$ (left boundary point) is not the root for any ϕ_is because of our choice of BCOPs, we add $x = 0$ as an additional root for all the ϕ_is.) other than $0, 1$ are given in Table 3.

$\phi_i \downarrow$	*roots* $\rightarrow$
ϕ_{10}	.00490 .04354 .11785 .22204 .34806 .48615 .62560 .75561 .86614 .94867
ϕ_{11}	.00413 .03677 .09991 .18943 .29947 .42284 .55146 .67691 .79103 .88629 .95661
ϕ_{14}	.00266 .02377 .06511 .12493 .20070 .28920 .38671 .48910 .59195 .69146 .78150 .86273 .92468 .97220

Table 3: Roots for the corresponding ϕ_is for test problem 2

Table 4 discuss the maximum error for the test problem 2 for various values of ϵ verses various ϕ_is. In this case too maximum error reduces as the value of ϵ becomes smaller.

Test Problem 3. (see [8]) We consider another interior layer SPP problem as

$$\epsilon u''(x) + xu'(x) = -\epsilon\pi^2\cos(\pi x) - \pi x\sin(\pi x), x \in (-1,1), 0 < \epsilon << 1, \tag{25}$$

with boundary conditions

$$u(-1) = -2,\ u(1) = 0. \tag{26}$$

$\epsilon \downarrow \phi_i \rightarrow$	ϕ_5	ϕ_{10}	ϕ_{11}	ϕ_{12}	ϕ_{13}	ϕ_{14}
0.1	0.2421	0.2333	0.2285	0.2238	0.2194	0.2152
0.01	0.3107	0.2730	0.2572	0.2463	0.2405	0 .2391
0.001	0.0520	0.1525	0 .1757	0 .1990	0 .2218	0.2435
0.0001	0.0055	0.0178	0.0210	0.0245	0.0283	0.0323
0.00001	0.00054	0.0018	0 .0021	0.0025	0.0029	0.0033

Table 4: Maximum error for various values of ϵ and mesh points for test problem 2

Its exact solution is not known and for small ϵ it gives a turning point near $x = 0$. Therefore we need finer mesh near the turning point to resolve the interior layer for small ϵ. To find the fitted mesh (finer mesh at $x = 0$) we define the inner product as

$$< f, g >= \int_{-1}^{1} \frac{f(x)g(x)}{\sqrt{1-x^2}} dx, \tag{27}$$

with weight function as $w(x) = 1/\sqrt{1-x^2}$ to have finer mesh near $x = 0$. In this case we choose f_is as

$$f_i = \{x, x^2, x^3, x^4, ...\} \tag{28}$$

and $h(x) = (x - 1)$. We can see that in this case $h(x)f_i$s do not satisfy the boundary conditions (i.e non zero b.c.) but at the point $x = 0$, where the interior layer occur, all the $h(x)f_i$s satisfy the given zero conditions. The roots of the few corresponding BCOPs other than 1 are given in Table 5 below.

$\phi_i \downarrow$	$roots \rightarrow$
ϕ_7	.20449 .56782 .84597 -.18876 -.55207 -.83014 -.98063
ϕ_8	.00620 .35035 .65300 .87768 -.33793 -.64057 -.86519 -.98471
ϕ_9	.16224 .46129 .71569 .90057 -.15219 -.45124 -.70563 -.89045 -.98762
ϕ_{10}	.00414 .28704 .54703 .76304 .91760 -.27875 -.53873 -.75473 -.90925 -.98977

Table 5: Roots for the corresponding ϕ_is for test problem 3

Figures 3 and 4 show the results for different values of ϵ and for $\epsilon = .000001$, we can see a thin interior layer at $x = 0$.

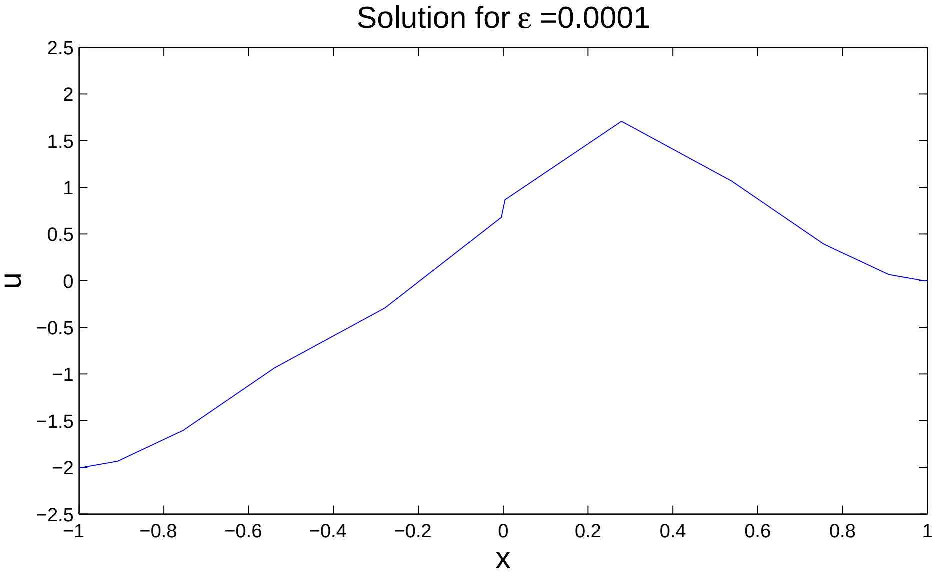


Figure 3: Approximate solution for $\epsilon = .0001$

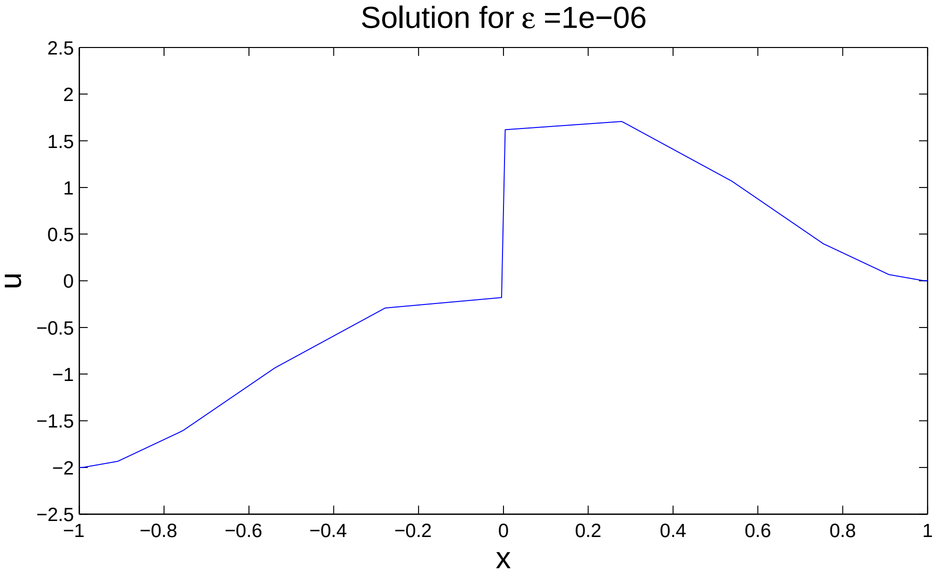


Figure 4: Approximate solution for $\epsilon = .000001$

Test Problem 4. (see [9]) We consider 2D linear reaction-diffusion problem as

$$-\epsilon^2 \triangle u(x,y) + 2u(x,y) = f(x,y), \text{ in } \Omega = (0,1) \times (0,1). \tag{29}$$

f has been chosen such that the exact solution of Eq. (29) is given as

$u(x,y) =$

$\phi_i \downarrow \epsilon \rightarrow$	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}	2^{-10}	2^{-15}
ϕ_8	0.00700	0.02057	0.01503	0.00424	0.00106	0.00026	0.00000026
ϕ_{10}	0.01570	0.00685	0.02144	0.00848	0.00217	0.00054	0.00000053

Table 6: Maximum error for various values of ϵ and mesh points for test problem 4

$$\left(1-\frac{\exp(-x/\epsilon)+\exp(-(1-x)/\epsilon)}{1+\exp(-1/\epsilon)}\right)\left(1-\frac{\exp(-y/\epsilon)+\exp(-(1-y)/\epsilon)}{1+\exp(-1/\epsilon)}\right).$$

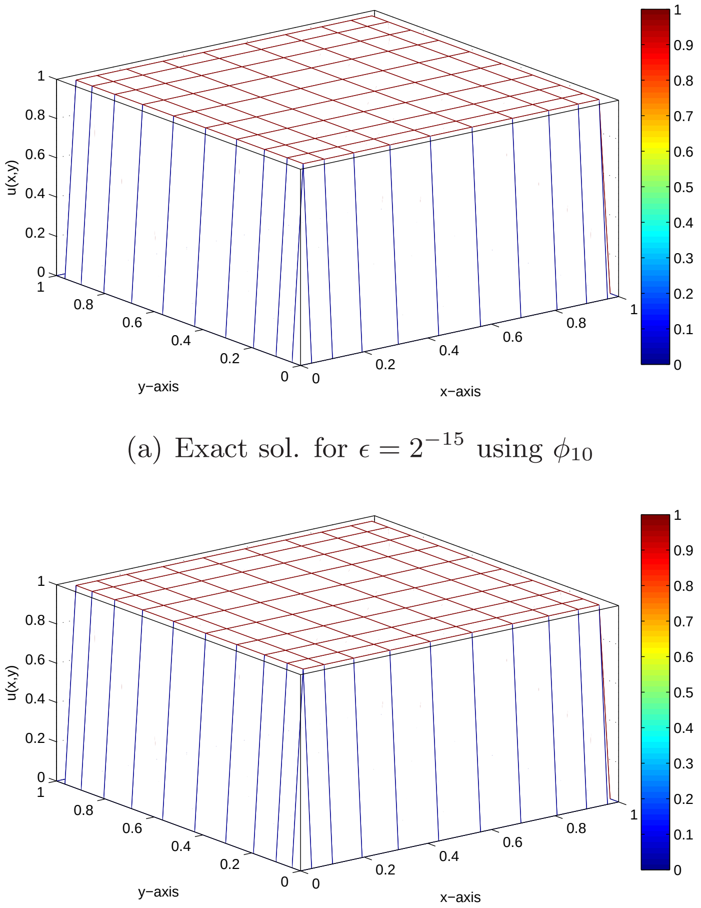
This $u(x,y)$ has typical boundary layers of width $O(\epsilon)$. Since the exact solution is known, we can accurately measure the maximum error as given in the Table 6. We have taken the same ϕ_is as discussed in example 1 (with zero boundary conditions). To solve 2D problem we have used the same 1D derivative matrices to generate Laplacian using tensor products, also known as *Kronecker products* as discussed in [10]. Table 6 gives the maximum errors, and Figure 5 shows the exact and computed results.

4. Conclusion

In this paper we have introduced the BCOPs methods to study the singularly perturbed problems having boundary and interior layers in one and two dimensions. The one of the advantage of BCOPs is that all the corresponding polynomials also satisfy the boundary conditions and there is no need to satisfy them separately. BCOPs have been used to solve problem in the vibration by treating them as the bases functions for the approximation. But in this paper we have taken the advantage of orthogonal collocation to develop fitted meshes to deal with SPPs.

Acknowledgement

Author would like to acknowledge Department of Science and Technology (DST), New Delhi, INDIA, for the financial grant under the fast track project SR/FTP/ MS-04/2008 and Prof. Bani Singh for all kind of help to learn and impliment BCOPs.



(a) Exact sol. for $\epsilon = 2^{-15}$ using ϕ_{10}

(b) Computed sol. for $\epsilon = 2^{-15}$ using ϕ_{10}

Figure 5: Solutions for example 4 (29)

References

[1] Bani Singh, S. Chakraverty, Boundary characteristic orthogonal polynomials in numerical approximation, *Communications in Numerical Methods in Engineering*, **10** (1994), 1027-1043.

[2] Bani Singh, S. Chakraverty, Flexural vibration of skew plates using boundary characteristic orthogonal polynomials in two variables, *Journal of Sound and Vibration*, **173** (1994), 158-178.

[3] Walter Gautshi, *Orthogonal Polynomials: Computation and Approximation*, Numerical Mathematics and Scientific Computation, Oxford University Press, USA (2004).

[4] J.J.H. Miller, E. O'Riordan, I.G. Shishkin, *Fitted Numerical Methods for Singular Perturbation Problems*, World Scientific (1996).

[5] P.A. Farrell, A.F. Hegarty, J.J.H. Miller, E. O'Riordan, I.G. Shishkin, *Robust Computational Techniques for Boundary Layers*, Chapman and Hall, CRC (2000).

[6] Vivek Kumar, High order compact finite difference scheme for sing. pert. reaction diffusion problems on a new mesh of Shishkin type, *J. of Optim. Th. and Appl.*, **143** (2009), 123-147.

[7] E.P. Doolan, J.J.H. Miller, W.H.A. Schilders, *Uniform Numerical Methods for Problems with Initial and Boundary Layers*, Boole Press, Dublin (1980).

[8] F. Mazzia, A. Sestini, D. Trigiante, The continuous extension of the B-spline linear multistep methods for BVPs on non-uniform meshes, *Applied Num. Math.*, **59** (2009), 723-738.

[9] J. Li, I.M. Navon, Uniformly converg. finite element methods for sing. pert. elliptic boundary value problems I: Reaction-diffusion type, *Comp. Math. Appl.*, **35**, No. 3 (1998), 57-70.

[10] R.A. Horn, C.R. Johnson, *Topics in Matrix Analysis*, Cambridge Univ. Press, Cambridge, UK (1991).

Comparison between classic PID, Integer Order PID and Fuzzy Logic Controller for Ceramic Infrared Heater: Analysis using MATLAB/Simulink

Vineet Shekher, Pankaj Rai, Om Prakash

Abstract*— This paper discusses the design, simulation and performance of ceramic infrared heater controller. This heater is energy saving potential, efficient heat transfer, uniform heating, efficient and instant heat. Many industries are increasibily making use of infrared technology as a means of improving their process. This type of heating often requires a large area of floor space. This study successfully developed a controller to achieve an effective and robust control of the infrared heating process. This paper consists three main tuning methods for IR heating system controller. Firstly, it presents design of PID controller using Zeigler Nichols (ZN) technique for first order plus time delay system using open loop step response method. Secondly, it presents the design of PID controller based on gain margin and phase margin (IOPID) for the same system. Thirdly, a fuzzy logic controller used for the same system for good stability and robust performance. Performance analysis shows the effectiveness of the ZN-PID, IOPID and fuzzy logic controller.

***Index Terms*— Zeigler Nichols, PID, IOPID, Gain margin, Phase margin, Fuzzy Logic**

I. Introduction

Conventional PID controller operates the majority of the control system in the world due to simple in algorithm, good in stability, high in reliability, easy in design and wide in adaption. The PID controller is used in wide range of problems like automotive, instrumentation, motor drives etc. PID controller provides robust and reliable performance for most of the systems if the PID parameters are tuned properly. Among these tuning methods the Zeigler Nichols (ZN) technique has been very influential. Zeigler Nichols presents two tuning methods, a step response method and ultimate frequency response method. In this paper we will investigate step response method for the IR heating controller.

In order to solve the problem, a control method which uses IOPID method based on gain and phase margin specification in temperature control for IR heating system is proposed in this paper. On the basis of gain and phase margin, IOPID controller can make entire use of the successful operations.

The field of fuzzy control has been making rapid progress in recent years .Fuzzy logic controller has been widely used for non-linear, time delay and high order system. The tuning of the parameters of the fuzzy module can be easily done by computational efforts. The methodology is shown to be effective for a higher static gain. In this paper, a novel methodology used based on the fuzzification. In this the value of the proportional gain is multiplied by a constant parameter less than one to reduce the overshoot, but has the drawback of increasing the rise time. To achieve both the aims of reducing the overshoot and decreasing the rise time, a fuzzy module depending upon the current output error and its derivative are used.

II. PID Tuning

The PID controller has the following standard form in the time domain [3]

$$C(t) = K_p \left[e(t) + T_d \frac{d}{dt} e(t) + \frac{1}{T_i} \int_0^t e(\tau) d\tau \right] \qquad (1)$$

where

$e(t)$ = system error;

K_p =Proportional gain;

T_d =Derivative gain constant;

T_i =Integral time constant;

We can also write equation (1) as

$$C(t) = K_p e(t) + K_d \frac{d}{dt} e(t) + K_i \int_0^t e(\tau) d\tau \qquad (2)$$

Where it is obviously $K_d = K_p T_d$ or $K_i = K_p / T_i$

The transfer function of PID controller is given as

$$C(s) = K_p + \frac{K_i}{s} + K_d \qquad (3)$$

Zeigler Nichols step response method is based on transient

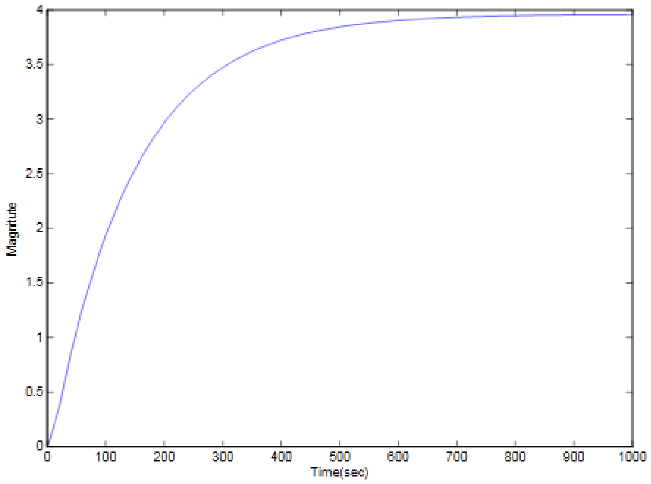
Manuscript received February 28, 2012.

Vineet Shekher, Electrical Engineering Department,DeenBandhu Chhotu Ram University science and Technology/Hindu College of Engineering/Sonepat/Haryana/India, 011-27931498/ 9034147386, (e-mail: vshekher2407@gamil.com).

Pankaj Rai, Electrical Engineering Department, Vinoba Bhave University, Birsa Institute of Technology, Sindri, Dhanbad, Jharkhand, India-mail: pr_bit2001@gmail.com).

Om Prakash, Chemical Engineering Department, Vinoba Bhave University, Birsa Institute of Technology, Sindri, Dhanbad, Jharkhand, India (e-mail:omprakash1151@gmail.com).

response experiment. The open loop step response of the IR heater under test resulted in the curve [1], [2] is shown in fig (1)



Fig(1): Mat Lab output of step response of IR heater Plant

The step response of IR heater is compared with the unit step response of a typical industrial process shown in fig (2) to determine the parameters of the process [3], [4].

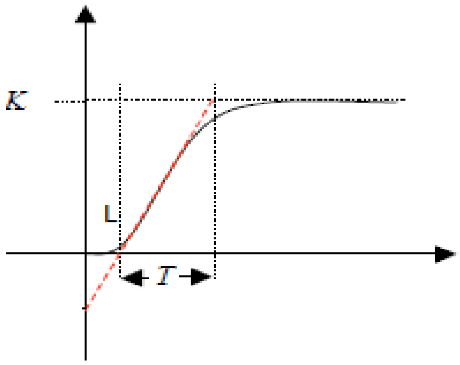


Fig (2): Step response with plant

From fig (2), the plant transfer function represented as

$$P(s) = \frac{K}{1+sT} e^{-sL} \tag{4}$$

Where K is the static gain, L is the time delay and T is the time constant

By comparing fig (1) and fig (2) the transfer function represented as

$$P(s) = \frac{3.96}{140s+1} e^{-7s} \tag{5}$$

Using the Zeigler – Nichols step response method formula in the table (1)

TABLE I: Tuning formula for ZN-PID

PID Controller	$K_{p=} \frac{1.2}{K}\left(\frac{T}{L}\right)$	$K_i = 2L$	$K_d = 0.5L$

From the general form for PID controller in equation (3)

$$C_{ZNPID}(s) = 6.06 + \frac{0.432}{s} + 21.21s \tag{6}$$

III. Design of Integer order PID controller

The open loop transfer function G(s) is

G(s) =C(s) P(s)

According to the first order plus time delay system transfer function P(s) we get its frequency response as

$$P(j\omega) = \frac{K}{jT\omega+1} e^{-jL\omega}$$

$$= \frac{K}{\sqrt{1+\omega^2T^2}} e^{-(\tan^{-1}(\omega T)+L\omega)}$$

The gain and phase of the plant are as follow

$$|P(j\omega)| = \frac{K}{\sqrt{1+\omega^2T^2}} \tag{7}$$

$$Arg[P(j\omega)] = -\tan^{-1}(\omega T) - L\omega \tag{8}$$

A. *Design Specification*

The objective of this paper is to design a fractional order controller so that the system fulfills different specifications regarding to the plant uncertainties, load disturbance and high frequency noise. Therefore, the design problem is formulated as follows [5],[10].

i. Phase Margin and Gain crossover frequency specification

Gain and phase margin have always served important parameter for robustness. It is known that the phase margin is related to the damping of the system. The equations that define the phase margin ϕ_{pm} and gain crossover frequency ω_{cp} are

$$|G(j\omega_{cp})| = |C(j\omega_{cp})P(j\omega_{cp})|_{dB} = 0dB \tag{9}$$

$$\text{Arg}[G(j\omega_{cp})] = \text{Arg}[C(j\omega_{cp})P(j\omega_{cp})] = -\pi + \phi_{pm} \tag{10}$$

ii. Robustness to gain variation in the gain of the plant

The gain variation of the plant demands that the phase directives w.r.t the frequency is zero, i.e. the phase bode plot is flat, at the gain crossover frequency.

$$\left(\frac{d(Arg(G(j\omega))}{d\omega}\right)_{\omega=\omega_{cp}} = 0 \tag{11}$$

B. Integer Order PID Controller Design

The open loop transfer function $G_1(s)$ for the IOPID with

FOPTD system is given as

$$G_1(s) = C(s)P(s) \qquad (12)$$

According to the IOPID controller transfer function (7), we can get the frequency response as,

$$C(j\omega) = K_p + \frac{K_i}{j\omega} + j\omega K_d$$

The gain and phase are as follow,

$$|C(j\omega)| = \sqrt{K_p{}^2 + (K_d\omega - (K_i/(\omega K_p))^2} \qquad (13)$$

$$Arg[C(j\omega)] = \tan^{-1}((K_d\omega^2 - K_i)/(\omega K_p)) \qquad (14)$$

Then the open loop frequency response

$$G_1(j\omega) = C(j\omega)P(j\omega)$$

The gain and phase of the open loop frequency response are as follows by using (5), (13) and (14)

$$|G_1(j\omega)| = |C(j\omega)||P(j\omega)| = \frac{K\sqrt{K_p{}^2 + (K_d\omega - (K_i/(\omega K_p))^2}}{\sqrt{1+\omega^2T^2}} \qquad (15)$$

$$Arg[G_1(j\omega)] = \tan^{-1}((K_d\omega^2 - K_i)/(\omega K_p)) - \tan^{-1}(\omega T) - L\omega \qquad (16)$$

According to specification (i), the phase of $G_1(j\omega)$ can be expressed as in as a form of ω_{cp} is

$$Arg[G_1(j\omega_{cp})] = \tan^{-1}((K_d\omega_{cp}{}^2 - K_i)/(\omega_{cp}K_p)) - \tan^{-1}(\omega_{cp}T) - L\omega_{cp}$$

$$= -\pi + \phi_{pm} \qquad (17)$$

Then,

$$\frac{K_d\omega^2{}_{cp} - K_i}{K_p\omega_{cp}} = A_1$$

Where $A_1 = \tan\left[\tan^{-1}(\omega_{cp}T) + L\omega_{cp} + \phi_{pm}\right]$

And according to specification (ii) about the robustness to gain variation in the plant,

$$\left(\frac{d(Arg(G_1(j\omega))}{d\omega}\right)_{\omega=\omega_{cp}} = 0$$

So

$$= \frac{d}{d\omega}\left(\tan^{-1}((K_d\omega_{cp}{}^2 - K_i)/(\omega_{cp}K_p)) - \tan^{-1}(\omega_{cp}T) - L\omega_c\right)_{\omega=\omega_{cp}} = 0 \qquad (18)$$

Then we get,

$$\frac{K_p(\omega_{cp}{}^2K_d + K_i)}{\omega_{cp}{}^2K_p{}^2 + (K_d\omega_{cp}{}^2 - K_i)} = \frac{T}{1+\omega_{cp}{}^2T^2} + L$$

Where

$$1 + \omega_{cp}{}^2T^2 = B_1$$

According to the specification (ii), we established an equation about K_p,

$$|G_1(j\omega_{cp})| = |C_1(j\omega_{cp})||P(j\omega_{cp})|$$

$$= \frac{\sqrt{K_p{}^2 + (K_d\omega_{cp} - (K_i/(\omega_{cp}K_p))^2}}{\sqrt{1+\omega_{cp}{}^2T^2}} = 1 \qquad (19)$$

From (17) (18) and (19), we can get

$$K_p = \frac{1}{k}\sqrt{\frac{B_1}{1+A_1{}^2}}$$

$$K_i = \frac{1}{2k}\left[\sqrt{\frac{1+A_1{}^2}{B_1}}(T\omega_{cp} + LB_1\omega_{cp}{}^2) - A_1\omega_{cp}\sqrt{\frac{B_1}{1+A_1{}^2}}\right]$$

$$K_d = \frac{1}{2k}\left[\sqrt{\frac{1+A_1{}^2}{B_1}}(T + LB_1) - A_1\omega_{cp}{}^{-1}\sqrt{\frac{B_1}{1+A_1{}^2}}\right]$$

If the parameter are set as follows

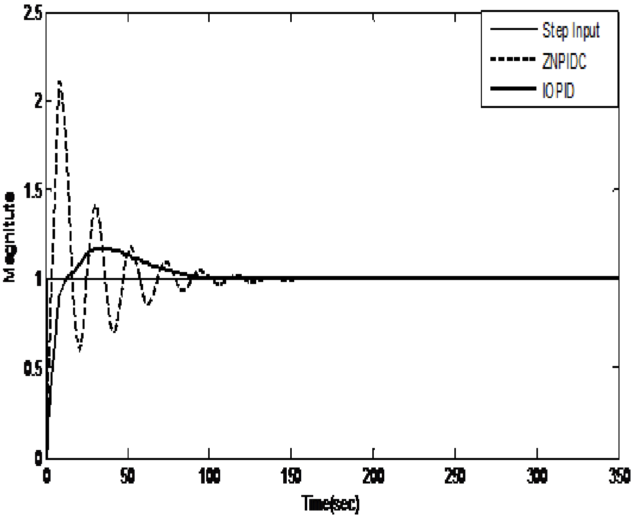
ω_{cp} =0.08 rad/sec, T=140sec, L=7sec, $\phi_{pm} = 60^\circ$

Then we get K_p, K_i and K_d directly

K_p =2.825, K_i =0.0855 and K_d =9.74

Let IOPID controller obtained from C_{IOPID} as

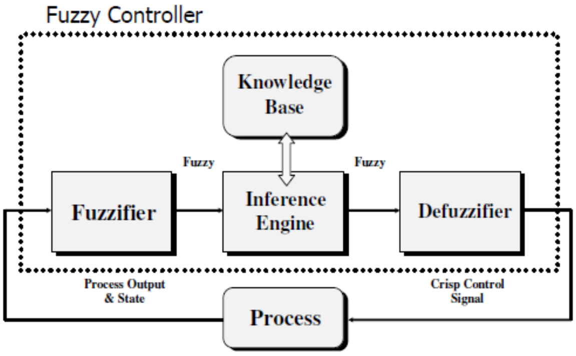
$$C_{IOPID} = 2.825 + \frac{0.0855}{s} + 9.74s \qquad (20)$$



Fig(3):- Step response of the system with ZNPIDC and IOPID

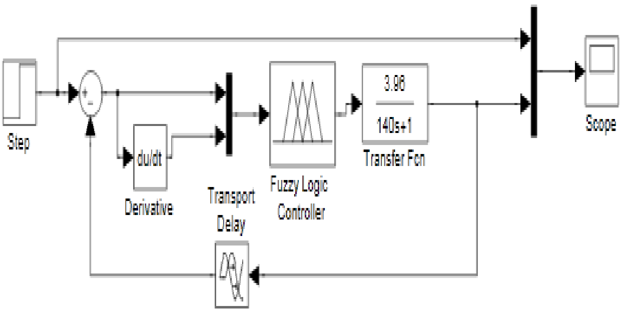
IV. FUZZY LOGIC CONTROLLER

Fuzzy logic control is based on fuzzy set theory, linguistic variable and fuzzy inference. Due to this fuzzy logic control is also been known as intelligent control[13] ,[14]. The basic block diagram of fuzzy logic control is shown in fig(4).



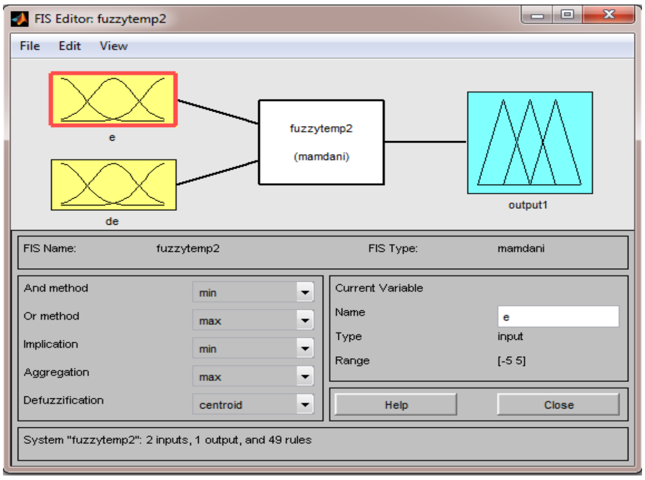
Fig(4):-Basic block diagram of fuzzy logic controller

In order to verify the effectiveness of the fuzzy logic controller and compare the control performance with the conventional PID controller and IOPID, we make a simulation with the below simplified model shown in fig(5) using Mat Lab simulink.



Fig(5):-Implementation of fuzzy logic controller with plant using simulink

For a two input fuzzy logic controller 3,5,7,9 or 11 membership function for each input are mostly uses. In this paper, only seven fuzzy membership function are used for two input error (e) and derivative error (de) and the fuzzy membership functions for the output parameter are shown in fig (6), here NB means negative big, NM means negative medium, NS means negative small, ZO means zero, PS means positive small, PM means positive medium and PB means positive big.



FIS Name: fuzzytemp2 FIS Type: mamdani

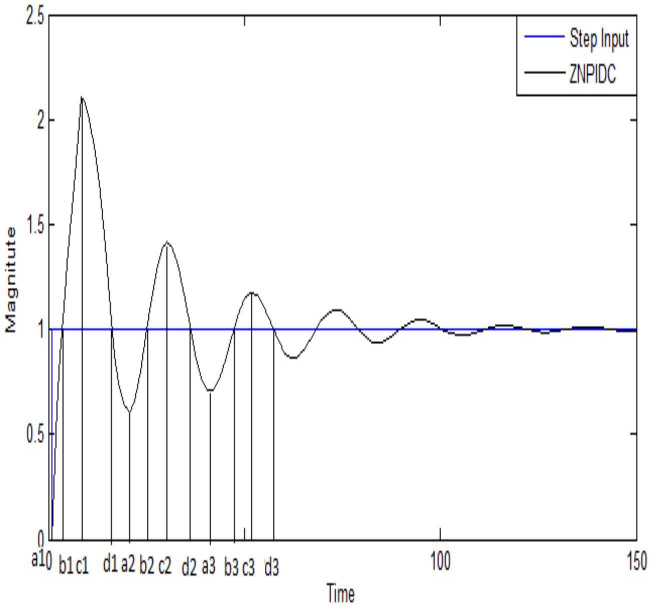
And method: min
Or method: max
Implication: min
Aggregation: max
Defuzzification: centroid

Current Variable
Name: e
Type: input
Range: [-5 5]

System "fuzzytemp2": 2 inputs, 1 output, and 49 rules

Fig(6): membership function editor for fuzzy controller

The fuzzy linguistic rules are defined from output response of the system using ZNPID controller as shown in fig(7)



Fig(7): Output response of plant with ZNPID controller for observation of deriving fuzzy control rules

This FLC design is quite intuitive and transparent to the users[15]. The rule base applies the appropriate control action depending on how far the response is moving towards the set point (i.e. error and derivative error).

Fig(7) illustrates the effectiveness of having direct control over the error and change of error in driving the temperature to a prescribed set point.

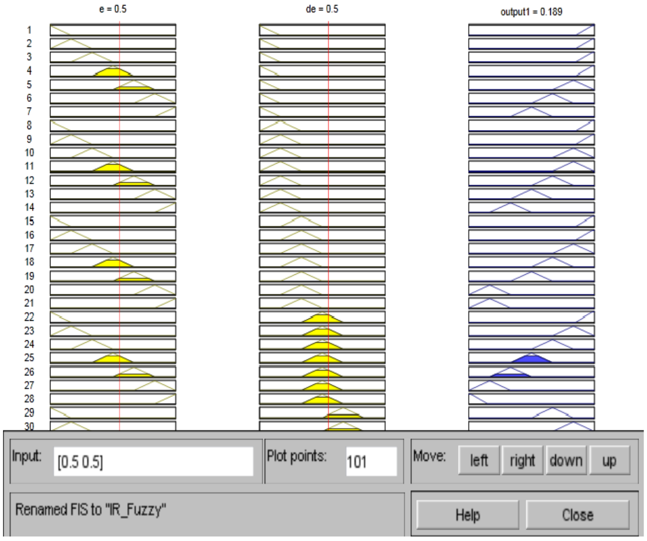
Table II :- Prototype of fuzzy control rules with term sets

Rule No.	e	de	du	Reference Point
1	PB	ZO	PB	a1
2	PM	ZO	PM	a 2
3	PS	ZO	PS	a 3
4	ZO	NB	NB	b 1
5	ZO	NM	NM	b 2
6	ZO	NS	NS	b 3
7	NB	ZO	NB	c 1
8	NM	ZO	NM	c 2
9	NS	ZO	NS	c 3
10	ZO	PB	PB	d 1
11	ZO	PM	PM	d 2
12	ZO	PS	PS	d 3
13	ZO	ZO	ZO	set point

The system response divided into seven phase. Depending upon the output is increasing or decreasing, 49 rules are been divided for the fuzzy logic controller shown in table (3).These 49 rules are sufficient to cover all possible situation for high static gain first order plus time delay system.

Table III: Rules base for fuzzy logic controller

de \ e	NB	NM	NS	ZO	PS	PM	PB
NB	PB	PB	PB	PB	PM	PS	ZO
NM	PB	PB	PM	PM	PS	ZO	NS
NS	PB	PB	PM	PS	ZO	NM	NM
ZO	PB	PM	PS	ZO	NS	NM	NB
PS	PS	PM	ZO	NS	NM	NB	NB
PM	PS	ZO	NS	NM	NM	NB	NB
PB	ZO	NS	NM	NB	NB	NB	NB



Input: [0.5 0.5]

Plot points: 101

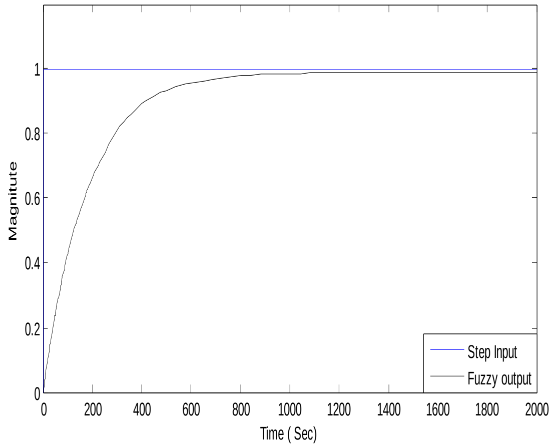
Move:

Renamed FIS to "IR_Fuzzy"

Fig(8):-Rule viewer for fuzzy logic controller

The Rule Viewer displays a roadmap of the whole fuzzy inference process and shows one calculation at a time. In this sense, it presents a sort of micro view of the fuzzy inference system.

Step response of the fuzzy logic based controlled first high order static gain first order plus time delay process is shown below



Fig(9): step response of fuzzy logic controller

V. Conclusion

This paper presented the design of PID controller using open loop step response method, Inter order PID controller based on gain margin and phase margin and fuzzy logic controller for high static gain based first order plus time delay system.

Simulation results using MatLab/Simulink are discussed for ZN tuned PID, IOPID and the Fuzzy logic controller. The Design is carried out in MATLAB and has been observed that the system response is improved after implementing fuzzy logic controller. The proposed methodology gives better performance in the rise time, peak overshoot and the steady state error. The responses observed from the IOPID controller have a slight over shoot which can be further improved by implementing fuzzy logic controller. But it is observed that, from the simulation that the fuzzy logic controller performs better response than IOPID controller and conventional PID controller.

REFERENCES

[1] Adonis, M and Khan, MTE. 2001. Infrared heating profile controller. *Proceedings of the 3rd International Conference on Control Theory and Applications*, Dec., 445-449.

[2] Adonis, M and Khan ‚MTE," PID control of infrared radiative power profile for ceramic emitters" , 2003 *IFAC*

[3] Astrom, K.J., and Hagglund, T.: 'Automatic tuning of PID controllers' (ISA, 1988)

[4] Ziegler, J.G., and Nichols, N.B.: 'Optimum settings for automatic controllers', *Trans.* ***ASME*** 1942, 64, pp. 759-768

[5] Ho, W.K., Hang, C.C,, and Cao, L.S.: 'Tuning of PID controllers based on gain and phase margin specifications', *Automatica*, 1995, 31, (3), pp. 497-502

[6] Astrom, K.J., and Hagglund, T.: 'PID controllers: theory, design, and tuning' (Instrument Society of America, 1995, 2nd edn.)

[7] R. S. Barbosa, J. A. Tenerio, Machado and Isabel. M. Ferreira, "Tuning of PID controllers based Bode's Ideal transfer function, *Nonlinear Daynamics*, vol. 38, pp.305- 321, 2004.

[8] D. Xue, Y.Q. Chen, D. P. Atherton "Linear Feedback Control Analysis and Design with MATLAB", Advances in Design and Control, Siam, 2007.

[9] Cvejn, J., 2009. Sub-optimal PID controller settings for FOPDT systems with long dead time. Journal of process control 19.

[10] Ho, W.K., Hang, C.C., Zhou, J.H., 1995. Performance and gain and phase margins of well-known PI tuning formulas.IEEE Transactions on Control Systems Technology 3.

[11] PID Controllers for Time-Delay Systems Guillermo J. Silva, Aniruddha Datta, S.R Bhattacharyya ,springer 2005

[12] C.H. Lee and C.C. Teng, "Tuning of PID Controllers for Stable and Unstable Processes based on gain and phase margin specifications", *International Journal of Fuzzy Systems, Vol. 3, No. 1,* pp. 346-355. 2001.

[13] Q. Yang, G. Li, X. Kang, Application of fuzzy PID control in the Heating System,Chinese Control and Descision Conference (CCDC2008).

[14] J. Wang, D. An, C. Lou, Application of fuzzy-PID controller in heating ventilating and air conditioning system, in: Proceedings of the IEEE International Conference on Mechatronics and Automation, China, 2006, pp. 2217–2222.

[15] Z.W. Woo, H.Y. Chung, J.J. Lin, A PID type fuzzy controller with self-tuning scaling factors, Fuzzy Sets and Systems 115 (2000) 321-326.

[16] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," Int. J. Man-Math. Stud., Vol. 7, pp. 1-13, 1975.

[17] G. K. I. Mann, B. G. Hu and R. G. Gosine, "Analysis of direct action fuzzy PID controller structures," IEEE Trans. SMC. – Pt. B, Vol. 29, pp. 371-388,Jun. 1999.

[18] Z.Y. Zhao, M. Tomizuka and S. Isaka, "Fuzzy gain scheduling of PID

controllers," IEEE Trans. Syst., Man, Cybern., Vol. 23, pp. 1392-1398,1993.

[19] S. G. Tzafestas, N. P. Papanikolopoulos, "Incremental fuzzy expert PID control," IEEE Trans. Ind. Electron., Vol. 37, No. 5, pp. 365-371, 1990.

[20] S. N. Sivanandam, S. Sumathi and S. N. Deepa,Introduction to Fuzzy Logic using MATLAB, Springer Berlin Heidelberg New York,2007.

Vineet Shekher received his B.E degree in Instrumentation from North Maharashtra University, Jalgoan, Maharashtra ,India in 1999, an M.E degree in Electrical Engineering (Control and Instrumentation) from Delhi College of Engineering ,Delhi University ,Delhi ,India in 2009.He is currently working toward the Ph.D. degree in Electrical Engineering ,Birsa Institute of Technology, Sindri, Dhanbad, Jharkhand, India. His research interests include linear control system, Non-linear control system and Process Instrumentation.

Dr. Pankaj Rai has obtained his B.Sc. Engg. Degree in Electrical Engg from MIT Muzaffarpur, Bihar in 1988 in first class with distinction, M.Tech. Degree in 2002 in Electrical Engg. (Control system) & Ph.D. degree in 2010 from Vinoba Bhave University, Hazaribagh, Jharkhand. His area of specialization is Fuzzy Logic & Neural Network application in Control and power system, presently working as head, This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity in an equation department of Electrical Engg., BIT Sindri, Dhanbad, Jharkhand, India.

Dr. Om Prakash obtained the degree of B.Sc. Engg. (Chemical) in 1973 from Ranchi University, completed M.Sc. Engg. (Chemical) specializing in petroleum Refinery Technology in 1982 from Ranchi University. He obtained his Ph. D. In 1996 from the department of Chemical Engineering, I. I. T. Delhi. After graduation he joined M/s KCI Ltd., Mirzapur as Apprentice Chemical Engineer, completed one year training in FCI (Fertilizer Corporation of India at Sindri unit and stared teaching career in 1979 joining B I T Sindri as teacher's fellow. Became Assistant Professor in 1983, subsequently promoted to Lecturer (Senior scale) in 1989 and became Assistant Professor in 1996 in the same college.

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Energy Procedia 14 (2012) 695 – 701

Energy Procedia

2nd International Conference on Advances in Energy Engineering (ICAEE 2011)

Comparison of Various Auxiliary Signals for Damping Subsynchronous Oscillations Using TCR-FC

Sanjiv Kumar[a*], Dr. Narendra Kumar[b] ,Vipin Jain[c]

[b]Department of Electrical Engineering,Delhi Technological University, Delhi-110042 India.
[a,c]Research Scholar, Delhi University, Delhi-110042 India.

Abstract

Because of large-capacity, long-distance and cross-region power transmission in our country's power grid, the subsynchronous oscillation (SSO) will arise from the use of series capacitor compensation and HVDC. Damping subsynchronous oscillation using static VAR compensators (SVC) is investigated in this paper.TCR-FC (Thyristor Controlled Rectifier with Fixed Capacitor) is a well known combination to improve voltage stability. Supplementary signals such as variation in reactive power, variation in frequency, variation in active power, variation in current can be used to enhance the dynamic response of the system. Their derivatives also can be used for better performance. In this paper three signals are compared and it is shown that Deviation in Reactive power gives best performance.

Keywords- FACTS; TCR-FC; Reactive power deviation; active power deviation; frequency deviation

1. Introduction

Control changing the network parameters is an effective method of improving transient stability. Flexible ac transmission system (FACTS) controllers due to their rapid response are suitable for transient stability control since they can bring about quick changes in the network parameters. Transient stability control involves changing the control variables such that the system state enters the stability region after a large disturbance [1].

Control by changing the network parameters is an effective method of improving transient stability. Flexible ac transmission system (FACTS) controllers due to their rapid response are suitable for transient stability control since they can bring about quick changes in the network parameters. Transient stability control involves changing the control variables such that the system state enters the stability region after a large disturbance.

* Corresponding author. Tel.: +91-9410436382
E-mail address: activesanjiv007@rediffmail.com

doi:10.1016/j.egypro.2011.12.887

TCR-FC (Thyristor Controlled Rectifier with Fixed Capacitor) is a well known combination to improve voltage stability. Supplementary signals are used to improve dynamics of power system, i.e to reduce power oscillations etc. Damping of power system oscillation plays an important role not only in increasing the transmission capability but also for stabilization of power system conditions after critical faults, particularly in weakly coupled networks. Series compensation has been widely used to enhance the power transfer capability. However, series compensation gives rise to dynamic instability and sub synchronous resonance (SSR) problems. Many preventive measures to cope with this dynamic instability problem in series compensated lines have been reported in literature.

These supplementary signals may be deviation in Reactive power, deviation in frequency, deviation in bus angle voltage, deviation in active power.etc. [3] [4]. Damping of power system oscillations plays an important role not only in increasing the power transmission capability but also for stabilization of power system conditions after critical faults [12]. In this paper deviation in reactive power, active power & frequency is used as Supplementary signal.

Subsynchronous resonance is addressed in three categories (i) induction generator effect (ii) torsional effect (iii) torque amplification. In all cases SSR is due to the interaction of a series capacitor with turbine generator. The first two types are caused by a steady state disturbance, while the third is excited by transient disturbance. Flexible AC transmission system (FACTS) technology provides unprecedented way for controlling transmission grids and increasing transmission capacity [7–9]. FACTS controllers have the flexibility of controlling both real and reactive power which could provide an excellent capability for improving power system dynamics. Several studies have investigated the potential of using this capability in mitigating SSR of series capacitive compensated transmission grids [10–15].

Two IEEE benchmark models have been proposed by the IEEE-SSR Working Group. These benchmark models have obtained world-wide acceptance and are extensively used for the study of different proposed damping devices SSR countermeasures [16-17].

The use of the thyristor controlled series capacitor (TCSC), static synchronous compensator and static synchronous series compensator in their balanced mode of operations has been implemented and/or studied as means for damping SSR. Generally FACTS controllers are used for power flow control and voltage stability [18-20]. Very less research has been carried out for damping of SSR using FACTS devices. In our research paper we have shown that UPFC is an effective FACTS device for damping of SSR. [21-28].

In this research paper we have developed a transmission system in MATLAB very similar to the IEEE first benchmark model.

It is helpful to look at the two SSR types within the classification which results from different sets of assumptions in our simplified model:

a) *Constant current field winding:* Only torsional interaction is present as a result of which currents and voltages at subsynchronous frequency in the stator and the shaft torque grow.

b) Damper winding on the rotor with zero initial current:

Only induction generator effect is present as a result of which currents and voltages at subsynchronous frequency in the stator, current in the rotor and the shaft torque grow.

c) Constant current field winding, constant synchronous speed: Interaction at subsynchronous frequency between the stator and rotor circuits stops. If there is no resistance in the stator, the subsynchronous currents and voltages resulting from initial conditions continue to exist without growing.

Also the electromagnetic torque component on the generator rotor is present, but since with our constant speed assumption we have actually placed an independent torque source on the generator rotor which completely compensates this torque, the mechanical oscillations do not grow.

d) Damper winding with zero initial current, constant synchronous speed: Interaction at synchronous frequency between the stator and rotor circuits stops. The electrical induction effect on the stator side results in currents and voltages at subsynchronous frequency to grow.

2. Study System

The study system consists of a steam turbine driven synchronous generator supplying bulk power to an infinite bus over a long transmission line. The study system, shown in Fig.1 consists of one synchronous generator, two transformers T_1 & T_2, one series capacitor and a TCR-FC in the middle of line. Modelling of above system & initial conditions are given in [1]. Block diagram of TCR-FC is shown in Fig.2. This block diagram includes firing control system and represented as a first order model having gain K and time constants T_1 and T_2.

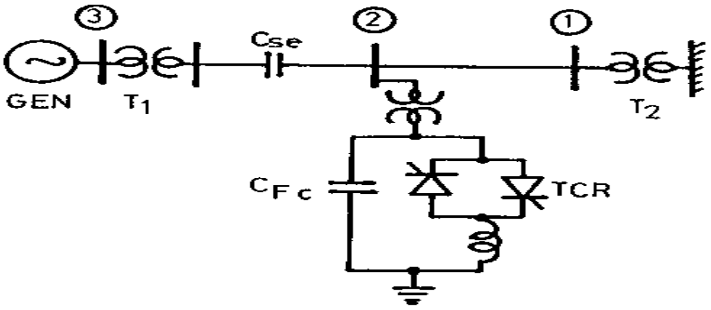


Fig.1

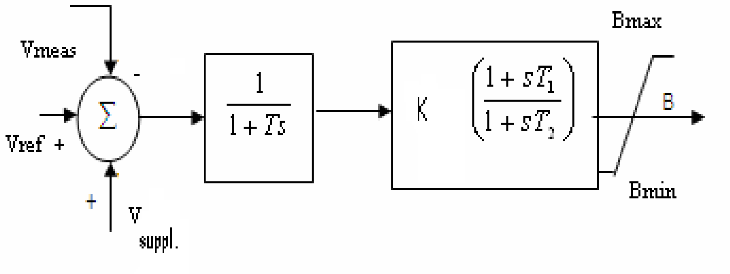


Fig.2

The controller send firing control signals to the thyristor switching unit to modify the equivalent susceptance of the TCR. In Fig. 2 V_{meas} is bus voltage at TCR –FC bus. V_{suppl} is change in voltage due to auxiliary signal deviation (reactive power, active power or frequency deviation) at TCR –FC bus.

3. Fault Simulation

In synchronous machine initial power is 2 pu. Suddenly it increased to 2.5 pu.

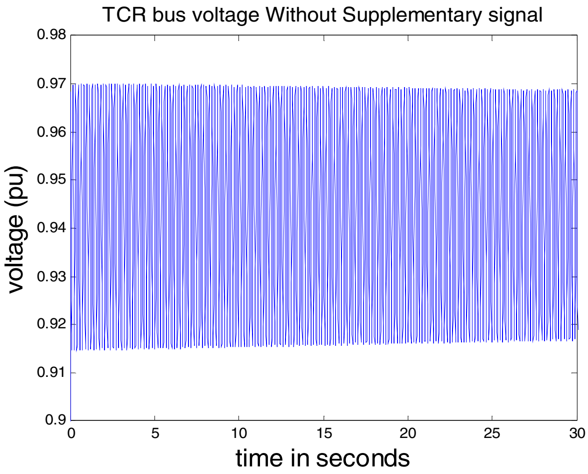


Fig.4

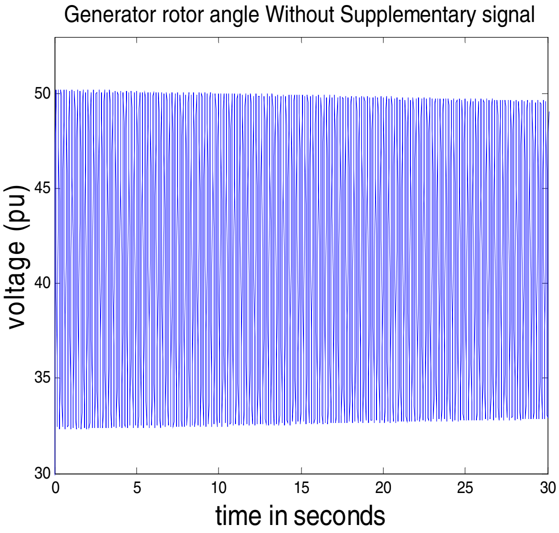


Fig. 5

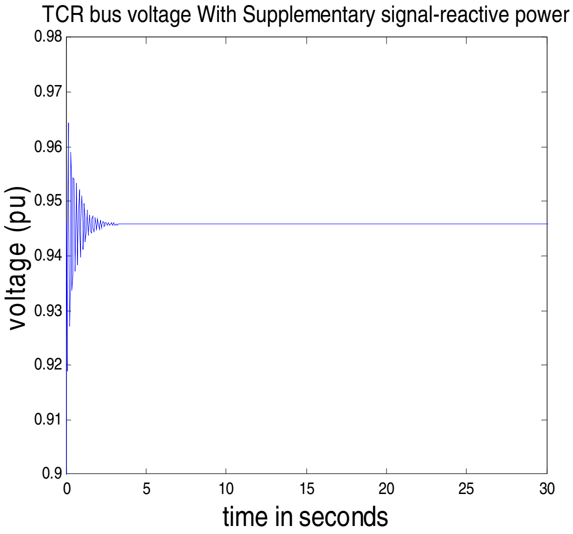


Fig. 6

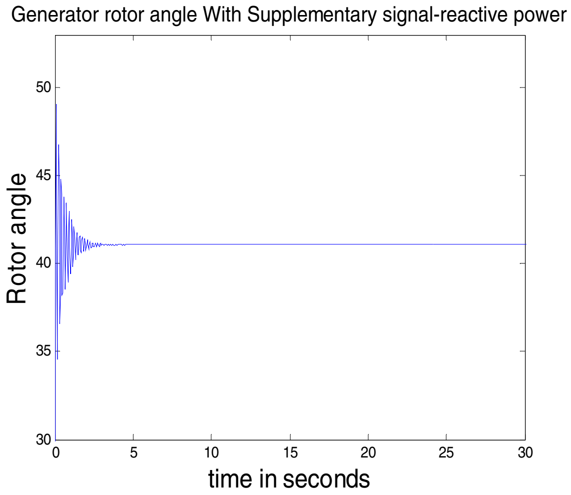


Fig. 7

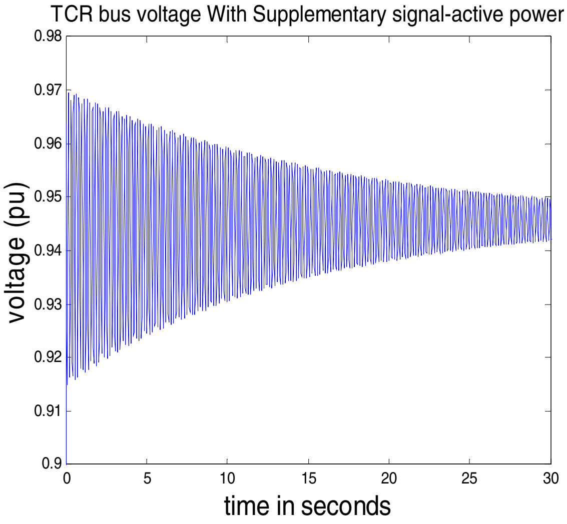


Fig. 8

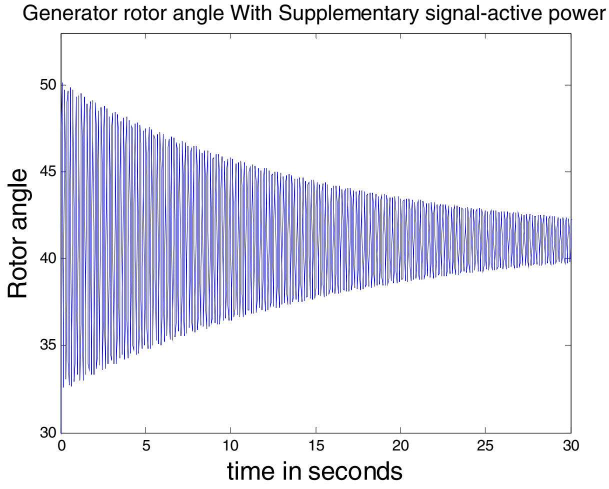


Fig. 9

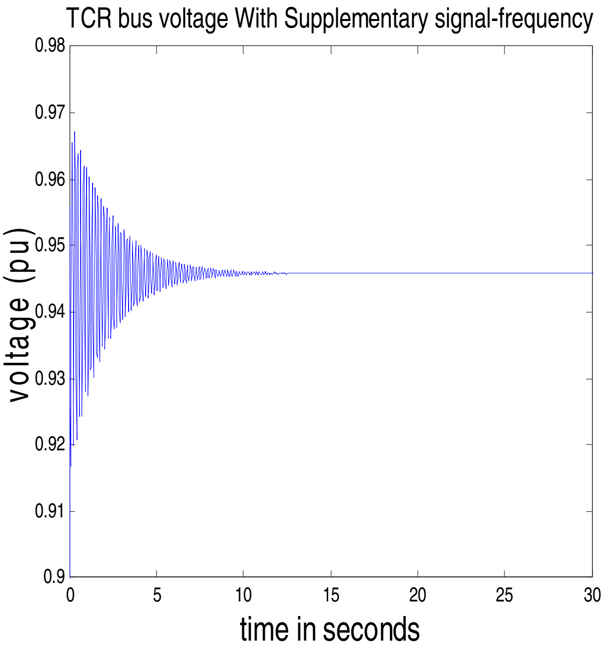


Fig. 10

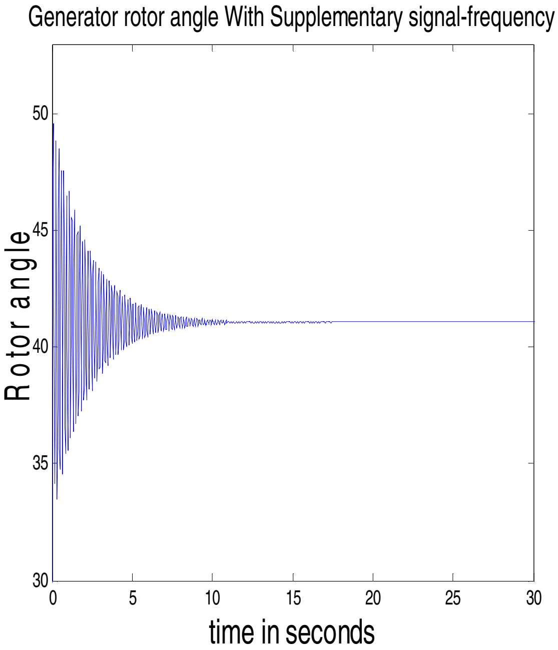


Fig.11

4. Results

(a) Without any supplementary controller - Voltage stability - Fig.4 Rotor stability - Fig.5
(b) With supplementary controller (auxiliary signal-reactive power) Voltage stability-Fig.6,Rotor stability Fig.7
(c) With supplementary controller (auxiliary signal-active power) Voltage stability -Fig. 8, Rotor stability Fig.9
(d) With supplementary controller (auxiliary signal-frequency deviation) Voltage stability -Fig. 10, Rotor stability Fig.11

5. Conclusion

In this paper the effectiveness of combined voltage and reactive power with TCR-FC auxiliary controllers have been evaluated for damping the subsynchronous oscillations in a given series compensated power system.Supplementary signal deviation in reactive power (Fig.6 & 7) gives best result, then frequency deviation gives little better results and out of three signals deviation in active power gives least result. No doubt that without Supplementary signal (Fig. 4 & 5) results are poorest. The response curves of terminal voltage, TCR-FC bus voltage, generator torque angle show a remarkable improvement and their oscillations die down effectively.The further work is going on with combining two signals in one controller.

Acknowledgements

The work presented in this paper has been performed under the AICTE R&D Project, “Enhancing the power system performance using FACTS devices” in the Flexible AC Transmission Research Laboratory at Delhi Technological University formerly Delhi College of Engineering, Delhi (India).

References

[1] N.Kumar, S.T.Nagarjan, "A SVS control strategy for damping Torsional Oscillation due to SSR in a series compensated system" Institution of Engineers (india), Vol. 91. March 2011

[2] Anderson,P.M,Agarwal,B.L.,Van Ness,J.E"Subsynchronous Resonance in Power System', IEEE Press.

[3] P.Kundur, Power System Stability and Control, Mc Graw Hill, 1994.

[4] Y. Wang, R.R. Mohler "Variable structure facts controller for power system transient stability" Transactions on Power Systems, Vol. 7, No. 1, February 1992

[5] W.Sae-Kok,A.Yokoyama,S.C.Verma " Excitation Control System Design of Rotary Type Frequency Converter for Performance Improvement of Power System Dynamics" IEEE Trans. on energy conversion, vol. 21, no. 1, March 2006

[6] M. Noroozian, L. Angquist, M. Ghandhari, G. Anderson" Improving power system dynamics by series-connected facts devices" IEEE Transactions on Power Delivery, Vol. 12, No. 4, October 1997.

[7] N.Karpagam, D.Devaraj "Application of GA for SVC –FACTS Controller for Power System Transient Stability Improvement" International Journal of Electrical Power and Energy Systems Engineering 2:2 2009

[8] Harbans Nakra, R. L.ewh Vaughan, Charles Gagnon-" Real-Time Simulator for Power System Dynamics Studies" IEEE Transactions on Power Systems, Vol. 10, No. 2, May 1995

[9] D. J. Trudnowski, M. K. Donnelly, J. F. Hauer –"Estimating Damping Effectiveness of BPA's Thyristor Controlled Series Capacitor by Applying Time and Frequency Domain Methods to Measured Response" IEEE Transactions on Power Systems, Vol. 11. No. 2, May 1996

[10] M. L. Crow , J. G. Chen"The multirate simulation of facts devices in power system dynamics" IEEE Transactions on Power Systems, Vol. 11, No. 1, February 1996

[11] Varma, R. K.; Auddy, S.; Semsedini, Y, "Mitigation of Subsynchronous Resonance in a Series-Compensated Wind Farm Using FACTS Controllers" IEEE Transaction on Power Delivery Page(s): 1645-1654 Vol.3,2008

[12] S.K. Gupta, Narendra Kumar, "Controlled Series Compensation in Coordination with Double Order SVS Auxiliary Controller and Induction Machine for repressing the Torsional Oscillations in Power system". Electric Power System Research 62, 93-103

Appendix A.

Generator data: 1110MVA, 22kV, R_a= 0.0036, X_L = 0.21
T_{do}'=6.66, T_{qo}'=0.44, T_{do}''=0.032, T_{qo}''=0.057s
X_d = 1.933, X_q = 1.743, X_d' =0.467, X_q' = 1.144, X_d'' = 0.312, X_q'' = 0.312 p.u.
IEEE type 1 excitation system:
T_R=0, T_A=0.02, T_E=1.0, T_F=1.0s, K_A=400, K_E=1.0; K_F=0.06 p.u.
V_{fmax}=3.9, V_{fmin}=0, V_{rmax}=7.3, V_{rmin}= -7.3
Transformer data:
R_T=0, X_T=0.15 p.u. (generator base)

Transmission line data:

Voltage 400kV, Length 600km, Resistance R=0.034Ω / km, Reactance X=0.325 Ω / km
Susceptance B_c=3.7μ mho / km

SVS data:

Six-pulse operation:

T_M=2.4, T_S=5, T_D = 1.667ms, K_I = 1200, K_P = 0.5, K_D = 0.01

Torsional spring-mass system data

Mass	shaft	Inertia H (s)	Spring constant K (p.u. torque/rad)

```
HP                          0.1033586
          HP-IP                                 25.772
IP                          0.1731106
          IP-LPA                                46.635
LPA                         0.9553691
          LPA-LPB                               69.478
LPB                         0.9837909
          LPB-GEN                              94.605
GEN                         0.9663006
         GEN-EXC                                 3.768
EXC                         0.0380697
-------------------------------------------------------------------------------------
```

Estimates for the Initial Coefficients of Bi-univalent Functions

[1]S. SIVAPRASAD KUMAR, [2]VIRENDRA KUMAR AND [3]V. RAVICHANDRAN
[1,2]Department of Applied Mathematics, Delhi Technological University, Delhi—110042, India
[3]School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia
[3]Department of Mathematics, University of Delhi, Delhi—110007, India
[1]spkumar@dce.ac.in, [2]vktmaths@yahoo.in, [3]vravi@maths.du.ac.in

Abstract. A bi-univalent function is a univalent function defined on the unit disk with its inverse also univalent on the unit disk. Estimates for the initial coefficients are obtained for bi-univalent functions belonging to certain classes defined by subordination and relevant connections with earlier results are pointed out.

2010 Mathematics Subject Classification: 30C45, 30C80

Keywords and phrases: Univalent functions, bi-univalent functions, coefficient estimate, subordination.

1. Introduction

Let $\mathscr{A}$ be the class of analytic functions defined on the open unit disk $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$ and normalized by the conditions $f(0) = 0$ and $f'(0) = 1$. A function $f \in \mathscr{A}$ has Taylor's series expansion of the form

$$f(z) = z + \sum_{n=2}^{\infty} a_n z^n. \tag{1.1}$$

An analytic function is *univalent* in a domain $D \subset \mathbb{C}$ if it is one-to-one in D. The class of all univalent functions in the open unit disk $\mathbb{D}$ of the form (1.1) is denoted by $\mathscr{S}$. Determination of the bounds for the coefficients a_n is an important problem in geometric function theory as they give information about the geometric properties of these functions. For example, the bound for the second coefficient a_2 of functions in $\mathscr{S}$ gives the growth and distortion bounds as well as covering theorems. Some coefficient related problems were investigated recently in [1, 3, 8, 9, 15, 17].

Since univalent functions are one-to-one, they are invertible and the inverse functions need not be defined on the entire unit disk $\mathbb{D}$. However, the famous Koebe one-quarter theorem ensures that the image of the unit disk $\mathbb{D}$ under every function $f \in \mathscr{S}$ contains a disk of radius $1/4$. Thus, the inverse of every function $f \in \mathscr{S}$ is defined on some disk containing the disk $|z| < 1/4$. It can also be easily verified that

$$f^{-1}(w) = w - a_2 w^2 + (2a_2^2 - a_3)w^3 - (5a_2^2 - 5a_2 a_3 + a_4)w^4 + \cdots \tag{1.2}$$

in some disk of radius at least 1/4. The function $f \in \mathscr{A}$ is *bi-univalent* in $\mathbb{D}$ if both f and f^{-1} are univalent in $\mathbb{D}$. In 1967, Lewin [14] introduced the class σ of bi-univalent analytic functions and showed that the second coefficient of every $f \in \sigma$ satisfy the inequality $|a_2| \leq 1.51$. Let σ_1 be the class of all functions $f = \phi \circ \psi^{-1}$ where ϕ, ψ map $\mathbb{D}$ onto domain containing $\mathbb{D}$ and $\phi'(0) = \psi'(0)$. In 1969, Suffridge [22] gave a function in $\sigma_1 \subset \sigma$ satisfying $a_2 = 4/3$ and conjectured that $|a_2| \leq 4/3$ for all functions in σ. Netanyahu [16] in 1969 proved this conjecture for the subclass of σ_1. Later in 1981, Styer and Wright [21] disproved the conjecture of Suffridge [22] by showing $a_2 > 4/3$ for some function in σ. See [4] also for an example to show $\sigma \neq \sigma_1$. For results on bi-univalent polynomial, see [13,18]. Branan [5] in 1967 conjectured that $|a_2| \leq \sqrt{2}$ for $f \in \sigma$. Kedzierawski [12, Theorem 2] in 1985 proved this conjecture for a special case when the function f and f^{-1} are starlike functions. The best known bound $|a_2| \leq 1.485$ is proved by Tan [23]. For some open problems and survey, see [11,19].

For $0 \leq \alpha < 1$, a function $f \in \sigma$ is in the class $\mathscr{S}^*_\sigma(\alpha)$ of *bi-starlike function of order* α, or $\mathscr{K}_\sigma(\alpha)$ of *bi-convex function of order* α if both f and f^{-1} are respectively starlike or convex functions of order α. For $0 < \alpha \leq 1$, the function $f \in \sigma$ is *strongly bi-starlike function of order* α if both the functions f and f^{-1} are strongly starlike functions of order α. The class of all such functions is denoted by $\mathscr{S}^*_{\sigma,\alpha}$. These classes were introduced by Branan and Taha [7] in 1985 (see also [6]). They obtained estimates on the initial coefficients a_2 and a_3 for functions in these classes. Recently, Ali *et al.* [2] extended the results of Branan and Taha [7] by generalizing their classes using subordination. For some related results for special cases, see [10, 20, 24]. Recall that an analytic function f is *subordinate* to an analytic function g, written $f \prec g$, if there is an analytic function w with $|w(z)| \leq |z|$ such that $f = g \circ w$. If g is univalent, then $f \prec g$ if and only if $f(0) = g(0)$ and $f(\mathbb{D}) \subseteq g(\mathbb{D})$. For the various applications of subordination one can refer to [1,3,9,15,17] and the references cited therein.

Throughout this paper, we assume that φ is an analytic univalent function with positive real part in $\mathbb{D}$, $\varphi(\mathbb{D})$ is symmetric with respect to the real axis and starlike with respect to $\varphi(0) = 1$, and $\varphi'(0) > 0$. The Taylor's series expansion of such function is of the form

$$\varphi(z) = 1 + B_1 z + B_2 z^2 + B_3 z^3 + \cdots \quad \text{with } B_1 > 0. \tag{1.3}$$

With this assumption on φ, we now introduce a class of functions investigated in the paper.

Definition 1.1. *Let $\lambda \geq 0$. A function $f \in \sigma$ given by (1.1) is in the class $\mathscr{R}_\sigma(\lambda, \varphi)$, if it satisfies*

$$(1-\lambda)\frac{f(z)}{z} + \lambda f'(z) \prec \varphi(z) \quad \textit{and} \quad (1-\lambda)\frac{F(w)}{w} + \lambda F'(w) \prec \varphi(w), \quad (F = f^{-1}).$$

The class $\mathscr{R}_\sigma(\lambda, \varphi)$ includes many earlier classes, which are mentioned below:

(1) $\mathscr{R}_\sigma(\lambda, (1+(1-2\beta)z)/(1-z)) = \mathscr{R}_\sigma(\lambda, \beta)$ $(\lambda \geq 1;\ 0 \leq \beta < 1)$ [10, Definition 3.1]
(2) $\mathscr{R}_\sigma(\lambda, ((1+z)/(1-z))^\alpha) = \mathscr{R}_{\sigma,\alpha}(\lambda)$ $(\lambda \geq 1;\ 0 < \alpha \leq 1)$ [10, Definition 2.1]
(3) $\mathscr{R}_\sigma(1, \varphi) = \mathscr{R}_\sigma(\varphi)$ [2, p. 345].
(4) $\mathscr{R}_\sigma(1, (1+(1-2\beta)z)/(1-z)) = \mathscr{R}_\sigma(\beta)$ $(0 \leq \beta < 1)$ [20, Definition 2]
(5) $\mathscr{R}_\sigma(1, ((1+z)/(1-z))^\alpha) = \mathscr{R}_{\sigma,\alpha}$ $(0 < \alpha \leq 1)$ [20, Definition 1]

Motivated by Ali *et al.* [2], we investigate the estimates for the initial coefficients a_2 and a_3 of bi-univalent functions belonging to the class $\mathscr{R}_\sigma(\lambda, \varphi)$ introduced above as well as to the classes $\mathscr{S}^\lambda_\sigma(\varphi)$ and $N^\lambda_{\sigma,\gamma}(\varphi)$ defined later. We also obtain an estimate for a_4 for functions belongs to $\mathscr{R}_\sigma(\lambda, \varphi)$. Our results generalize several well-known results in [2, 10, 20] and these are pointed out.

2. Coefficient estimates

Our first result provides estimates for the coefficients a_2, a_3 and a_4 for functions belonging to the class $\mathscr{R}_\sigma(\lambda, \varphi)$.

Theorem 2.1. *If $f \in \mathscr{R}_\sigma(\lambda,\varphi)$, then*

$$|a_2| \le \frac{B_1\sqrt{B_1}}{\sqrt{|(1+2\lambda)B_1^2+(1+\lambda)^2(B_1-B_2)|}}, \tag{2.1}$$

$$|a_3| \le \frac{B_1^2}{(1+\lambda)^2}+\frac{B_1}{1+2\lambda} \tag{2.2}$$

and

$$|a_4| \le \frac{3B_1+2|B_2|+|B_3|}{1+3\lambda}. \tag{2.3}$$

Proof. Since $f \in \mathscr{R}_\sigma(\lambda,\varphi)$, there exists two analytic functions $r,s:\mathbb{D}\to\mathbb{D}$, with $r(0)=0=s(0)$, such that

$$(1-\lambda)\frac{f(z)}{z}+\lambda f'(z)=\varphi(r(z)) \text{ and } (1-\lambda)\frac{F(w)}{w}+\lambda F'(w)=\varphi(s(z)). \tag{2.4}$$

Define the functions p and q by

$$p(z)=\frac{1+r(z)}{1-r(z)}=1+p_1z+p_2z^2+p_3z^3+\cdots \text{ and } q(z)=\frac{1+s(z)}{1-s(z)}=1+q_1z+q_2z^2+q_3z^3+\cdots, \tag{2.5}$$

or equivalently,

$$r(z)=\frac{p(z)-1}{p(z)+1}=\frac{1}{2}\left(p_1z+\left(p_2-\frac{p_1^2}{2}\right)z^2+\left(p_3+\frac{p_1}{2}\left(\frac{p_1^2}{2}-p_2\right)-\frac{p_1p_2}{2}\right)z^3+\cdots\right) \tag{2.6}$$

and

$$s(z)=\frac{q(z)-1}{q(z)+1}=\frac{1}{2}\left(q_1z+\left(q_2-\frac{q_1^2}{2}\right)z^2+\left(q_3+\frac{q_1}{2}\left(\frac{q_1^2}{2}-q_2\right)-\frac{q_1q_2}{2}\right)z^3+\cdots\right). \tag{2.7}$$

It is clear that p and q are analytic in $\mathbb{D}$ and $p(0)=1=q(0)$. Also p and q have positive real part in $\mathbb{D}$, and hence $|p_i|\le 2$ and $|q_i|\le 2$. In the view of (2.4), (2.6) and (2.7), clearly

$$(1-\lambda)\frac{f(z)}{z}+\lambda f'(z)=\varphi\left(\frac{p(z)-1}{p(z)+1}\right) \text{ and } (1-\lambda)\frac{F(w)}{w}+\lambda F'(w)=\varphi\left(\frac{q(w)-1}{q(w)+1}\right). \tag{2.8}$$

Using (2.6) and (2.7) together with (1.3), it is evident that

$$\begin{aligned}\varphi\left(\frac{p(z)-1}{p(z)+1}\right) &= 1+\frac{1}{2}B_1p_1z+\left(\frac{1}{2}B_1\left(p_2-\frac{1}{2}p_1^2\right)+\frac{1}{4}B_2p_1^2\right)z^2\\ &\quad+\left(\frac{B_1}{2}\left(2p_3+p_1\left(\frac{p_1^2}{2}-p_2\right)-p_1p_2\right)+\frac{B_2p_1}{2}\left(p_2-\frac{p_1^2}{2}\right)+\frac{B_3p_1^3}{8}\right)z^3+\cdots.\end{aligned} \tag{2.9}$$

and

$$\begin{aligned}\varphi\left(\frac{q(w)-1}{q(w)+1}\right) &= 1+\frac{1}{2}B_1q_1w+\left(\frac{1}{2}B_1\left(q_2-\frac{1}{2}q_1^2\right)+\frac{1}{4}B_2q_1^2\right)w^2\\ &\quad+\left(\frac{B_1}{2}\left(2q_3+q_1\left(\frac{q_1^2}{2}-q_2\right)-q_1q_2\right)+\frac{B_2q_1}{2}\left(q_2-\frac{q_1^2}{2}\right)+\frac{B_3q_1^3}{8}\right)w^3+\cdots.\end{aligned} \tag{2.10}$$

Since $f \in \sigma$ has the Maclaurin series given by (1.1), a computation shows that its inverse $F = f^{-1}$ has the expansion given by (1.2). It follows from (2.8), (2.9) and (2.10) that

$$(1+\lambda)a_2 = \frac{1}{2}B_1p_1, \tag{2.11}$$

$$(1+2\lambda)a_3 = \frac{1}{2}B_1\Big(p_2 - \frac{1}{2}p_1^2\Big) + \frac{1}{4}B_2p_1^2, \tag{2.12}$$

$$(1+3\lambda)a_4 = \frac{B_1}{2}\left(2p_3 + p_1\Big(\frac{p_1^2}{2} - p_2\Big) - p_1p_2\right) + \frac{B_2p_1}{2}\left(p_2 - \frac{p_1^2}{2}\right) + \frac{B_3p_1^3}{8}, \tag{2.13}$$

$$-(1+\lambda)a_2 = \frac{1}{2}B_1q_1, \tag{2.14}$$

$$(1+2\lambda)(2a_2^2 - a_3) = \frac{1}{2}B_1\left(q_2 - \frac{1}{2}q_1^2\right) + \frac{1}{4}B_2q_1^2 \tag{2.15}$$

and

$$\begin{aligned} -(1+3\lambda)(5a_2^2 - 5a_2a_3 + a_4) = \frac{B_1}{2}&\left(2q_3 + q_1\Big(\frac{q_1^2}{2} - q_2\Big) - q_1q_2\right) \\ &+ \frac{B_2q_1}{2}\left(q_2 - \frac{q_1^2}{2}\right) + \frac{B_3q_1^3}{8}. \end{aligned} \tag{2.16}$$

From (2.11) and (2.14), it follows that

$$p_1 = -q_1 \tag{2.17}$$

and

$$8(1+2\lambda)^2a_2^2 = B_1^2(p_1^2 + q_1^2). \tag{2.18}$$

Now (2.12), (2.15) and (2.18) yield

$$a_2^2 = \frac{(p_2+q_2)B_1^3}{4[(1+2\lambda)B_1^2 + (1+\lambda)^2(B_1 - B_2)]}. \tag{2.19}$$

Thus the desired estimate on $|a_2|$ as asserted in (2.1), follows at once using the fact that $|p_2| \leq 2$ and $|q_2| \leq 2$.

By subtracting (2.12) from (2.15) and a computation using (2.19) and (2.11) finally lead to

$$a_3 = \frac{B_1^2p_1^2}{4(1+\lambda)^2} + \frac{(q_2 - p_2)B_1}{4(1+2\lambda)},$$

which in turn yields the estimate given in (2.2).

By adding (2.13) and (2.16) and a computation using (2.19) leads to

$$-5(1+3\lambda)(a_2^2 - a_2a_3) = \frac{B_1}{2}(p_3+q_3) - \frac{B_1}{2}(p_1p_2 - q_1q_2) + \frac{B_2}{2}(p_1p_2 + q_1q_2). \tag{2.20}$$

Now subtracting (2.16) from (2.13), will yield

$$\begin{aligned} 5(1+3\lambda)(a_2^2 - a_2a_3) + 2(1+3\lambda)a_4 = B_1(p_3 - q_3) &+ \frac{B_1}{4}(-2p_1p_2 + 2q_1q_2 + p_1^3) \\ &+ \frac{B_2}{2}(2p_1p_2 + 2q_1q_2 - p_1^3) + \frac{B_3p_1^3}{4}. \end{aligned} \tag{2.21}$$

Again the equations (2.20) and (2.21) lead to

$$2(1+3\lambda)a_4 = B_1p_3 + \frac{B_1}{4}(-4p_1p_2+p_1^3) - \frac{B_2}{2}(2p_1p_2-p_1^3) + \frac{B_3p_1^3}{4}. \tag{2.22}$$

Now the equation (2.22) can be rewritten as

$$\begin{aligned} 2(1+3\lambda)a_4 &= B_1p_3 + \frac{B_1}{4}\left(-2p_1p_2 + 2p_1\left(\frac{p_1^2}{2} - p_2\right)\right) \\ &\quad - B_2p_1\left(p_2 - \frac{p_1^2}{2}\right) + \frac{B_3p_1^3}{4}. \end{aligned} \tag{2.23}$$

Finally an application of the known result,

$$|p_i| \le 2 \quad \text{and} \quad \left|p_2 - \frac{p_1^2}{2}\right| \le 2 - \frac{|p_1|^2}{2} \le 2$$

in (2.23), yields the desired estimate given by (2.3) for a_4. ▮

Remark 2.1. For $\lambda = 1$, Theorem 2.1 reduces to a result of Ali *et al.* [2, Theorem 2.1].

For $\varphi(z) = (1+Cz)/(1+Dz)$, $-1 \le D < C \le 1$, Theorem 2.1 leads to the following result:

Corollary 2.1. *Let* $-1 \le D < C \le 1$. *If* $f \in \mathscr{R}_\sigma(\lambda, (1+Cz)/(1+Dz))$, *then*

$$|a_2| \le \frac{C-D}{\sqrt{(1+2\lambda)(C-D)+(1+\lambda)^2(1+D)}},$$

$$|a_3| \le \frac{(C-D)^2}{(1+\lambda)^2} + \frac{C-D}{1+2\lambda}, \quad \textit{and} \quad |a_4| \le \frac{(C-D)(3+2|D|+D^2)}{1+3\lambda}.$$

For $C = 1-2\beta$ with $0 \le \beta < 1$ and $D = -1$, Corollary 2.1 reduces to the following result [10, Theorem 3.1], as well as it gives an estimate for $|a_4|$.

Example 2.1. Let $0 \le \beta < 1$ and $\lambda \ge 0$. If $f \in \mathscr{R}_\sigma(\lambda, \beta)$, then

$$|a_2| \le \sqrt{\frac{2(1-\beta)}{1+2\lambda}}, \quad |a_3| \le \frac{4(1-\beta)^2}{(1+\lambda)^2} + \frac{2(1-\beta)}{1+2\lambda}, \quad \text{and} \quad |a_4| \le \frac{12(1-\beta)}{1+3\lambda}.$$

For $\lambda = 1$ and $\varphi(z) = (1+z)/(1-z)$, Theorem 2.1 gives the following coefficient estimates for $f \in \mathscr{R}_\sigma(0)$:

$$|a_2| \le \sqrt{\frac{2}{3}} \approx 0.816, \ |a_3| \le \frac{5}{3} \approx 1.667 \text{ and } |a_4| \le 3.$$

Since the estimate on $|a_2|$ for $f \in \mathscr{R}_\sigma(0)$ is improved over the conjectured estimate $|a_2| \le \sqrt{2} \approx 1.414$ for $f \in \sigma$, the functions in $\mathscr{R}_\sigma(0)$ are not the candidate for the sharpness of the estimate in σ.

When $\varphi(z) = ((1+z)/(1-z))^\alpha$, $0 < \alpha \le 1$ in Theorem 2.1, we get the following corollary. The estimates for a_2 and a_3 is the same as [10, Theorem 2.1] while the estimate for $|a_4|$ is new.

Corollary 2.2. *Let* $0 < \alpha \le 1$ *and* $\lambda \ge 0$. *If* $f \in \mathscr{R}_\sigma(\lambda, \alpha)$, *then*

$$|a_2| \le \frac{2\alpha}{\sqrt{(1+\lambda)^2 + \alpha(1+2\lambda-\lambda^2)}},$$

$$|a_3| \leq \frac{4\alpha^2}{(1+\lambda)^2} + \frac{2\alpha}{1+2\lambda}, \quad \textit{and} \quad |a_4| \leq \frac{4\alpha(\alpha^2+3\alpha+5)}{3(1+3\lambda)}.$$

Definition 2.1. *Let $\lambda \geq 0$. A function $f \in \sigma$ is in the class $\mathscr{S}^{\lambda}_{\sigma}(\varphi)$, if it satisfies*

$$\left(\frac{f(z)}{z}\right)^{\lambda-1} f'(z) \prec \varphi(z) \quad \textit{and} \quad \left(\frac{F(w)}{w}\right)^{\lambda-1} F'(w) \prec \varphi(w) \quad (F = f^{-1}).$$

Note that for a suitable choice of λ and φ, the class $\mathscr{S}^{\lambda}_{\sigma}(\varphi)$, reduces to the following known classes:

(1) $\mathscr{S}^0_{\sigma}((1+(1-2\beta)z)/(1-z)) = \mathscr{S}^*_{\sigma}(\beta) \quad (0 \leq \beta < 1)$.
(2) $\mathscr{S}^0_{\sigma}(((1+z)/(1-z))^{\alpha}) = \mathscr{S}^*_{\sigma,\alpha} \quad (0 < \alpha \leq 1)$.
(3) $\mathscr{S}^1_{\sigma}((1+(1-2\beta)z)/(1-z)) = \mathscr{R}_{\sigma}(\beta) \quad (0 \leq \beta < 1)$.
(4) $\mathscr{S}^1_{\sigma}(((1+z)/(1-z))^{\alpha}) = \mathscr{R}^*_{\sigma,\alpha} \quad (0 < \alpha \leq 1)$.

Theorem 2.2. *If $f \in \mathscr{S}^{\lambda}_{\sigma}(\varphi)$, then*

$$|a_2| \leq \frac{\sqrt{2B_1}B_1}{\sqrt{|(\lambda^2+3\lambda+2)B_1^2+2(\lambda+1)^2(B_1-B_2)|}} \tag{2.24}$$

and

$$|a_3| \leq \frac{2(B_1+|B_2-B_1|)}{\lambda^2+3\lambda+2}. \tag{2.25}$$

Proof. Since $f \in \mathscr{S}^{\lambda}_{\sigma}(\varphi)$, there are analytic functions $r,s : \mathbb{D} \to \mathbb{D}$, with $r(0) = s(0) = 0$, satisfying

$$\left(\frac{f(z)}{z}\right)^{\lambda-1} f'(z) = \varphi(r(z)) \text{ and } \left(\frac{F(w)}{w}\right)^{\lambda-1} F'(w) = \varphi(s(z)). \tag{2.26}$$

Let p and q be defined as in (2.5), then it is clear from (2.26), (2.6) and (2.7) that

$$\left(\frac{f(z)}{z}\right)^{\lambda-1} f'(z) = \varphi\left(\frac{p(z)-1}{p(z)+1}\right) \text{ and } \left(\frac{F(w)}{w}\right)^{\lambda-1} F'(w) = \varphi\left(\frac{q(z)-1}{q(z)+1}\right). \tag{2.27}$$

It follows from (2.27), (2.9) and (2.10) that

$$(1+\lambda)a_2 = \frac{1}{2}B_1p_1, \tag{2.28}$$

$$-\frac{(1-\lambda)(\lambda+2)}{2}a_2^2 + (\lambda+2)a_3 = \frac{1}{2}B_1\left(p_2 - \frac{1}{2}p_1^2\right) + \frac{1}{4}B_2p_1^2, \tag{2.29}$$

$$-(1+\lambda)a_2 = \frac{1}{2}B_1q_1 \tag{2.30}$$

and

$$\frac{(\lambda+2)(\lambda+3)}{2}a_2^2 - (\lambda+2)a_3 = \frac{1}{2}B_1\left(q_2 - \frac{1}{2}q_1^2\right) + \frac{1}{4}B_2q_1^2. \tag{2.31}$$

The equations (2.28) and (2.30) yield

$$p_1 = -q_1 \tag{2.32}$$

and

$$2(1+\lambda)^2 a_2 = \frac{1}{4}B_1^2(p_1^2+q_1^2). \tag{2.33}$$

From (2.29), (2.31), (2.32) and (2.33), it follows that

$$a_2^2 = \frac{B_1^3(p_1^2+q_1^2)}{2[(\lambda^2+3\lambda+2)B_1^2+2(B_1-B_2)(\lambda+1)^2]}, \tag{2.34}$$

which yields the desired estimate on $|a_2|$ as described in (2.24). Similarly, it can be obtained from (2.29), (2.31) and (2.32) that

$$a_3 = \frac{B_1(p_2(\lambda+3)+q_2(1-\lambda))+2(B_2-B_1)p_1^2}{4(\lambda^2+3\lambda+2)},$$

which eventually leads to the desired estimate (2.25) on a_3. ▮

Remark 2.2. If $\lambda = 0$, then the Theorem 2.2 reduces to [2, Corollary 2.1] and when $\varphi(z) = (1+(1-2\beta)z)/(1-z)$ $(0 \leq \beta < 1)$, it reduces to [7, Theorem 3.1] .

Definition 2.2. *Let $0 \neq \gamma \in \mathbb{C}$ and $\lambda \geq 0$. A function f given by (1.1) is said to be in the class $N_{\sigma,\gamma}^{\lambda}(\varphi)$, if f and $F = f^{-1}$ satisfy the subordinations*

$$1+\frac{1}{\gamma}\left(\frac{zf'(z)+\lambda z^2 f''(z)}{\lambda z f'(z)+(1-\lambda)f(z)}-1\right) \prec \varphi(z)$$

and

$$1+\frac{1}{\gamma}\left(\frac{wF'(w)+\lambda w^2 F''(w)}{\lambda w F'(w)+(1-\lambda)F(w)}-1\right) \prec \varphi(w).$$

Note that by choosing appropriate values for λ and γ, the class $N_{\gamma}^{\lambda}(\varphi)$ reduces to different classes:

(1) $N_{\sigma,1}^{0}((1+(1-2\beta)z)/(1-z))) = \mathscr{S}_{\sigma}^{*}(\beta) \quad (0 \leq \beta < 1)$.
(2) $N_{\sigma,1}^{1}(((1+(1-2\beta)z)/(1-z))) = \mathscr{K}_{\sigma}(\beta) \quad (0 \leq \beta < 1)$.
(3) $N_{\sigma,1}^{0}(((1+z)/(1-z))^{\delta}) = \mathscr{S}_{\sigma,\delta}^{*} \quad (0 < \delta \leq 1)$.

Theorem 2.3. *If $f \in N_{\sigma,\gamma}^{\lambda}(\varphi)$, then*

$$|a_2| \leq \frac{|\gamma| B_1\sqrt{B_1}}{\sqrt{|(1+2\lambda-\lambda^2)B_1^2\gamma^2+(1+\lambda)^2(B_1-B_2)|}} \quad \text{and} \quad |a_3| \leq \frac{|\gamma|(B_1+|B_2-B_1|)}{|1+2\lambda-\lambda^2|}.$$

The proof is omitted as it is similar to the proof of Theorem 2.2.

Remark 2.3. If we set $\gamma = 1$ and $\varphi(z) = (1+(1-2\beta)z)/(1-z)$ $(0 \leq \beta < 1)$ in Theorem 2.3, then for $\lambda = 0$ and $\lambda = 1$, it respectively reduces to [7, Theorem 3.1] and [7, Theorem 4.1].

References

[1] R. M. Ali, N. E. Cho, N. Jain and V. Ravichandran, Radii of starlikeness and convexity of functions defined by subordination with fixed second coefficients, Filomat, accepted

[2] R. M. Ali, S. K. Lee, V. Ravichandran and S. Supramaniam, Cofficient estimates for bi-univalent function Ma-Minda starlike and convex functions, Appl. Math. Lett., **25** (2012), 344–351.

[3] R. M. Ali, S. Nagpal and V. Ravichandran, Second-order differential subordination for analytic functions with fixed initial coefficient, Bull. Malays. Math. Sci. Soc. (2) **34** (2011), no. 3, 611–629.

[4] D. Bshouty, W. Hengartner and G. Schober, Estimates for the Koebe constant and the second coefficient for some classes of univalent functions, Canad. J. Math. **32** (1980), no. 6, 1311–1324.

[5] A. Brannan and J. G. Clunie, Aspects of contemporary complex analysis Proceedings of the NATO Advanced Study Institute held at the University of Durham, Durham, July 120, 1979, Academic Press New York, London, 1980.

[6] D. A. Brannan and T. S. Taha, On some classes of bi-univalent functions, in *Mathematical analysis and its applications (Kuwait, 1985)*, 53–60, KFAS Proc. Ser., 3 Pergamon, Oxford.

[7] D. A. Brannan and T. S. Taha, On some classes of bi-univalent functions, Studia Univ. Babeş-Bolyai Math. **31** (1986), no. 2, 70–77.

[8] Sh. Chen, S. Ponnusamy and X. Wang, Coefficient estimates and Landau-Bloch's constant for planar harmonic mappings, Bull. Malays. Math. Sci. Soc. (2) **34** (2011), no. 2, 255–265.

[9] N. E. Cho and O. S. Kwon, A class of integral operators preserving subordination and superordination, Bull. Malays. Math. Sci. Soc. (2) **33** (2010), no. 3, 429–437.

[10] B. A. Frasin and M. K. Aouf, New subclasses of bi-univalent functions, Appl. Math. Lett. **24** (2011), no. 9, 1569–1573.

[11] A. W. Goodman, An invitation to the study of univalent and multivalent functions, Internat. J. Math. Math. Sci. **2** (1979), no. 2, 163–186.

[12] A. W. Kedzierawski, Some remarks on bi-univalent functions, Ann. Univ. Mariae Curie-Skłodowska Sect. A **39** (1985), 77–81 (1988).

[13] A. Kedzierawski and J. Waniurski, Bi-univalent polynomials of small degree, Complex Variables Theory Appl. **10** (1988), no. 2-3, 97–100.

[14] M. Lewin, On a coefficient problem for bi-univalent functions, Proc. Amer. Math. Soc. **18** (1967), 63–68.

[15] J.-L. Liu, Certain sufficient conditions for strongly starlike functions associated with an integral operator, Bull. Malays. Math. Sci. Soc. (2) **34** (2011), no. 1, 21–30.

[16] E. Netanyahu, The minimal distance of the image boundary from the origin and the second coefficient of a univalent function in $|z| < 1$, Arch. Rational Mech. Anal. **32** (1969), 100–112.

[17] S. Supramaniam, R.M. Ali, S. K. Lee and V. Ravichandran, Convolution and differential subordination for multivalent functions, Bull. Malays. Math. Sci. Soc. (2) **32** (2009), no. 3, 351–360.

[18] H. V. Smith, Bi-univalent polynomials, Simon Stevin **50** (1976/77), no. 2, 115–122.

[19] H. V. Smith, Some results/open questions in the theory of bi-univalent functions, J. Inst. Math. Comput. Sci. Math. Ser. **7** (1994), no. 3, 185–195.

[20] H. M. Srivastava, A. K. Mishra and P. Gochhayat, Certain subclasses of analytic and bi-univalent functions, Appl. Math. Lett. **23** (2010), no. 10, 1188–1192.

[21] D. Styer and J. Wright, Result on bi-univalent functions, Proc. Amer. Math. Soc., Vol.82, No 2, 1981, 243–248.

[22] T. J. Suffridge, A coefficient problem for a class of univalent functions, Michigan Math. J. 16(1969), 33-42.

[23] D. L. Tan, Coefficient estimates for bi-univalent functions, Chinese Ann. Math. Ser. A **5** (1984), no. 5, 559–568.

[24] Q.-H. Xu, Y.-C. Gui and H. M. Srivastava, Coefficient estimates for a certain subclass of analytic and bi-univalent functions, Appl. Math. Lett. **25** (2012), 990–994.

Identification and Proof of Ownership by Watermarking Relational Databases

Vidhi Khanduja, *Member, IACSIT* and O. P. Verma

***Abstract*—Rapid increase in copying and distributing digital assets are major concerned to content owners. In this paper, we proposed a new robust secure and imperceptible embedding mechanism to resolve the two important concerns namely; owner identification and proof of ownership. The steps of proposed mechanism for watermarking relational databases mainly involves encoding and decoding on numerical attribute of relational database in three phases; 1)Watermark preparator, 2)Watermark position detector and 3) Watermark Embedder or Detector. The first phase resolves ownership identification issue as owner's identity is used to get watermark bits. In second phase position where watermarks are to be embedded are identified using secret key and pseudorandom generators. This phase marks multiple attributes with varying number of candidate bit positions within a single tuple. In the third phase watermarks are embedded in Encoder. While decoder extracts watermarks and detects database piracy.**

***Index Terms*—Relational Database, Watermark, Copyright protection, Ownership identification, Proof of ownership.**

I. Introduction

Internet is an excellent distribution system for digital media because it is inexpensive, eliminates warehousing and stock and delivery is almost instantaneous. Copying and distributing digital assets have become layman's task. However, owners of such digital assets are concerned about the copyright of their products.

The general solution to this problem is watermarking. A watermark is information that can be used for ownership verification and proof of identity of owner of digital products. Watermarking techniques allows owner of data to embed an imperceptible watermark into data which can include anything the owner chooses. Watermark embedding for relational data is made possible by the fact that real data can tolerate a small amount of error without any significant degradation in their usability [1]. There are many application contexts for which data represent an important asset, ownership of which must be carefully enforced. Any watermarking system should satisfy following properties:

1. *Embedding effectiveness*: The probability that the embedder will successfully embed a watermark in a randomly selected database.
2. *Fidelity*: The perceptual quality of watermarked content.
3. *Blind detection*: Detecting watermark should not require original database.
4. *Robustness*: The ability of watermark to survive normal processing of content
5. *Security*: The ability of the watermark to resist hostile attacks.
6. *Modification and multiple watermarks:* The possibility of changing embedded watermarks or embedding several watermarks in one tuple of the database.

In the proposed method all above properties are taken in to the consideration.

II. Related Work

Zhi-Hao Zhang, Xiao-Ming jin, Jain-Min wan [2] proposed image-based novel watermarking method for the numeric data. In their method an identification image is embedded into relational data for representing copyright information. Several other image-based watermarking mechanisms [3]-[6] are proposed in literature for watermarking numeric and non-numeric attributes. However [7], [8] proposed different mechanism for watermarking relational databases based on partitioning the databases and then embedding watermarks into them. C.Jiang, X.Chen, Zhi Li [9] proposed the watermarking algorithm, which can embed the watermark into relational database in DWT domain**.** D.Hanyurwimfura, Y.Liu and Z.Liu [10] watermarks non-numeric multi words data based on lavenshtein distance. H.Cui, X.Cui, M.Meng [11] proposed a public key cryptography based algorithm for watermarking relational databases.

The watermarking algorithm for relational databases proposed in [1] assume that database relations can be watermarked in some attributes, such that changes in few values do not affect their applications. This algorithm embeds watermarks only in one attribute out of several candidates attributes in a tuple.

In this paper we proposed a technique to securely and randomly select any number of attributes out of selected candidate attributes for embedding watermarks in varying number of least significant bits. We have devised a secure and imperceptible embedding mechanism that provides not only proof of ownership but also owner identification.

III. Proposed Algorithm

Proposed watermarking system consists of two subsystems watermark encoder and respective decoder.

Watermark Encoder: It embeds desired watermarks into relational database. This task is achieved using three steps as shown in Fig.1.

Manuscript received January 20, 2012; revised March 3, 2012.

Vidhi Khanduja was with department of Information Technology, Delhi Technological University. She is currently pursuing her PhD from Delhi University, India (e-mail: vidhikhanduja9@gmail.com).

O. P. Verma is with the Information Technology Department, Delhi Technological University, Delhi, India (e-mail: opverma.dce@gmail.com).

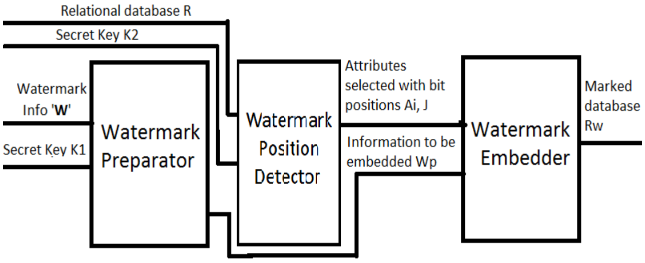


Fig. 1. Watermark Encoder

A. *Watermark Preparator*

Watermark to be inserted is selected by owner of the database. The watermark must be chosen such that it reflects owner's identity. This step identifies the identity of database's copyright holder as watermark. Thus ensures owners' identification.

Owner selects the watermarking text 'W' and secret key 'K1' to create a watermark to be embedded.

The algorithm

```
1.  Input the values of 'W' and K1
2.  For each character $C_i$ in W do
3.  $W_b[i]=C_i + K1$
    [end of for loop]
4.  $W_P$=binary ($W_b$)  // binary($W_b$) function
                          converts number to binary.
```

Line 2 in the algorithm indicates that owner chosen text is read character by character and addition of each character with secret key is computed in line 3 to give integer value. These values are stored in W_b array. At line 4, binary of W_b is taken and finally stored in W_p array.

B. *Watermarking Position Detector*

Suppose R is relation whose scheme is R $(P, A_0, A_1,...., A_{n-1})$ where P is primary key attribute and R contains total n attributes. Let owner selects 'v' number of numeric attributes that are candidates for marking. Each attribute A_i is numeric with values such that small changes in LB_{Ai} least significant bits are imperceptible. We consider that each attribute has varying number of candidate bit positions i.e. $LB_{Ai.}$ The gap γ [1] is a control parameter that determines the number w of tuples marked out of total r tuples via approximate relationship $w= r / \gamma$. The t.X represents the value of attribute X in tuple t ε R.

In this algorithm cryptographic pseudorandom sequence generators (CPSG) [12] are used that generates computationally infeasible sequence of numbers which depends on initial seed. Pseudorandom generators generate same fixed sequence of numbers every time if fixed initial seed is given.

The following functions are used in the algorithm

1) MAC: For each tuple't' in relation R, secure Message Authentication Code[13] is computed using secret key K2 known only to owner of the database and tuple's primary key t.P.
2) Next(CPSG1): This generates next number in random sequence using CPSG1.
3) Selectattr(next(CPSG1)): An another pseudorandom sequence generator CPSG2 is created with initial seed as next(CPSG1) whose output is a vector with number of states equivalent to v. These states decide what all attributes in a tuple are selected for watermark. Since output of this depends on previous pseudorandom generator, this increases the level of security.

For erasing a watermark, the attacker needs to correctly guess the tuples that are marked and the selected attributes with their corresponding selected bit positions.

The algorithm

```
1.  Input the value of secret key K2.
2.  For each tuple t ε R do
3.  Compute MAC = H(K2|| t.P || K2)
        Where, H( ) is secure hash function,
        and ' ||' is concatenation operator.
4.  Seed CPSG1 with MAC of each tuple.
5.  If (next(CPSG1) mod γ equals 0)  then
                                  //mark the tuple
6.  Attrindc[ ]= selectattr(next(CPSG1))
7.  For each value in Attrindc[ ]
8.  If (Attrindc[i] equals 1)   // mark the attribute
9.  Select $A_i$ for marking
10. Bitindex j=next(CPSG1) mod $LB_{Ai}$
                  // mark corresponding bit position
    [end of if at line 8]
    [end of for loop at line 7]
    [end of if at line 5]
    [end of for loop at line 2]
```

C. *Embed Watermark*

For selected attribute A_i and corresponding selected bit position j, we embed watermark generated W_p in relational database R. If number of watermark bits in W_p are less then number of detected watermarked positions in step2 we repeat the watermark bits in W_p again.

Watermark Decoder:

Fig. 2 shows watermark decoder which detects whether the database is pirated or not.

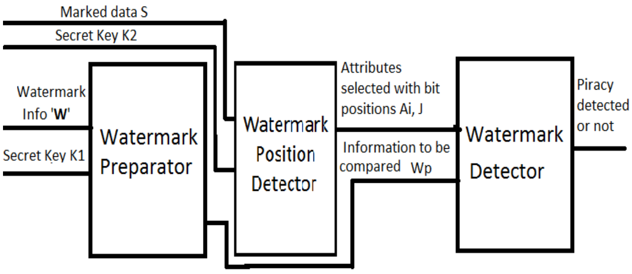


Fig. 2. Watermark Decoder

In detection process, the first two steps of watermark insertion are followed. Once attribute indices and bit positions are found in marked database S using secret key K2, we test whether or not the bits value matches the values that should have been assigned by insertion algorithm and count the number of matches matchcnt(m) against total number of watermarks totalcount(w). If there are very many matches or very few matches we suspect piracy [1]. We fix small value α ϵ (0, 1) and sets

$$\tau = \max\{\, t\square \left[0, \frac{w}{2}\right] : \sum_{i=t}^{w-t} b\left(i; w, \frac{1}{2}\right) \geq 1 - \alpha\} \qquad (1)$$

where

$$b(i; n, p) = nkp^{i}(1-p)^{n-i}$$

We suspects piracy if either m < τ τ or m > w-τ τ, as probability of so few or so many matches under null hypothesis is less than or equal to α. α is called significance level of the test.

Functions used in watermark detector algorithm:

1) Match(s. Ai, j): This function test whether or not the bit value of attribute s. Ai at position j matches the values that is assigned by embedding algorithm i.e Wp and returns 1 if match found.
2) Threshold(totalcount, α): This function calculates threshold value τ using (1). Total number of watermarks inserted and value of α are passed to the function.

The algorithm

```
//Watermark Preparation
 1. Input the values of watermark information 'W' and
    secret key K1
 2. For each character $C_i$ in W do
 3.   $W_b[i]=C_i + K1$
    [end of for loop]
 4. $W_P$=binary ($W_b$)   // binary($W_b$) function
                           converts number to binary.

//Watermark Position Detection
5.  Input the value of secret key K2.
6.  Totalcount=matchcnt=0
7.  For each tuple t ε S do
8.  Compute MAC = H(K2|| t.P || K2)
        where, H( ) is secure hash function,
        and ' ||' is concatenation operator.
9.  Seed CPSG1 with MAC of each tuple.
10. If (next(CPSG1) mod γ equals 0)  then
                                    //mark the tuple
11. Attrindc[ ]= selectattr(next(CPSG1))
12. For each value in Attrindc[ ]
13. If (Attrindc[i] equals 1)   // mark the attribute
14. Select $A_i$ for marking
15. Bitindex j=next(CPSG1) mod $LB_{Ai}$
                      // mark corresponding bit position
16. totalcount=totalcount+1

// Watermark Detector
17. matchcnt=matchcnt+match(s.$A_i$,j)
18. τ  = threshold(totalcount, α )
19. If ((matchcnt< τ) or (matchcnt>totalcount-τ))
    then
20. Suspect piracy
     [end of if at line 19]
     [end of if at line 13]
     [end of for loop at line 12]
     [end of if  at line 10]
    [end of for loop of line 7]
```

Detecting watermark is blind technique as it does not require original database and watermarks can be detected even in small subset of watermark relations as long as sample contains some of the marks (discussed in Section IV).

For ownership identity, the watermark bits are extracted from database S and reverse of the watermark preparation algorithm is followed to get repeated watermarked text from which original W is extracted.

III. Experiments and Analysis

The proposed algorithm is tested and evaluated on an experimental database consisting of approximately 10000 tuples. We ran the experiment on MATLAB environment and found that our algorithm is robust against following types of attacks.

A. *Subset Deletion Attack*

In this, attacker may delete randomly selected subset of tuples of watermarked database so that watermark will be removed.

We performed the experiment by deleting selected subsets of database and watermark extracted was recorded as shown in Fig. 3. Our experiment revealed that even if 90% of subsets are deleted approximately 12% of watermarks are still detected. Thus it still provides as a proof of ownership and to great extent ownership identification as watermarking bits are repeatedly embedded, we can extract meaningful information by further processing.

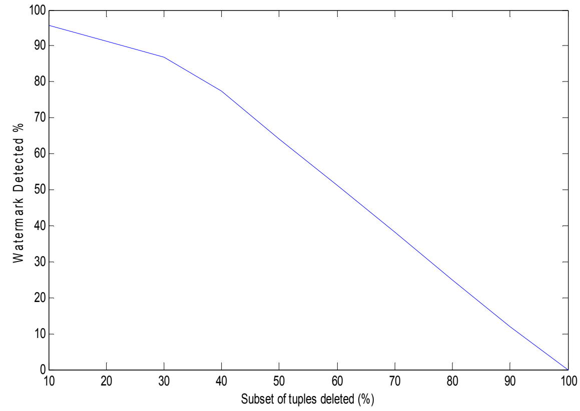


Fig. 3. Watermark Detection in Subset deletion attack.

B. *Subset Addition attack*

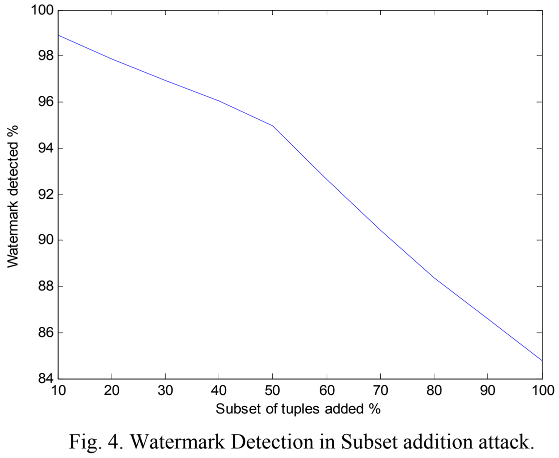


Fig. 4. Watermark Detection in Subset addition attack.

In this attacker may add subset of tuples to watermarked database so that watermark will be removed.

We performed the experiment by adding selected subsets to database and watermark extracted was recorded as shown in Fig.4. Our experiment revealed that this attack has very little impact on extraction of watermarked database. Ownership identification information is extracted completely.

IV. Conclusion

Owner identification and proof of ownership issues are resolved in this paper. This paper proposes a secure robust and imperceptible algorithm. We divide embedding algorithm in three phases: Watermark preparator, watermark position detector and watermark embedder. The ownership identification issue is resolved by embedding owner's identity as watermark in Preparator phase. Position detector phase securely identifies multiple attributes with varying number of candidate bit positions of the single table. Embedder inserts watermarks at identified bit positions of multiple attributes of relational database. The robustness of the proposed algorithm was verified against number of database attacks.

References

[1] R. Agrawal, Peter J. Haas, J.Kiernan, "Watermarking relational data: framework, algorithms and analysis," *The VLDB Journal*, pp. 157-169, 2003.

[2] Z.-H. Zhang, X.-M. Jin, J.-M. Wan, "Watermarking relational database using image," in *IEEE proc. Of Third International Conference on Machine Learning and Cybernetics*, 2004, pp. 1739-1744.

[3] A. Al-Haj and A. Odeh, "Robust and blind watermarking of relational database systems," *Journal of Computer Science* vol. 4, no. 12, pp. 1024-1029, 2008.

[4] J. Sun, Z. Cao, and Z. Hu," Multiple watermarking relational databases using image", in *IEEE proc. of International Conference on MultiMedia and Information Technology*, 2008, pp. 373-376.

[5] Z. Hu, Z. Cao, and J. Sun," An image based algorithm for watermarking relational databases", in *IEEE proc. International Conference on Measuring Technology and Mechatronics Automation,* 2009, pp. 425-428.

[6] A. Odeh and A. Al-Haj, "Watermarking relational database systems," in *IEEE proc. First International Conference on the Applications of Digital Information and Web Technologies ICADIWT* , 2008, pp. 270-274.

[7] A. Deshpande and J. Gadge," New watermarking technique for relational databases," in *proc. of IEEE ICETET*, 2009, pp. 664-669.

[8] S. Bhattacharya and A. Cortesi,"A distortion free watermark framework for relational databases," *in proc. ICSOFT (2),* 2009, pp. 229-234.

[9] C. Jiang, X. Chen, and Z. Li "Watermarking relational databases for ownership protection based on DWT," in *proc. Fifth International Conference on Information Assurance and Security*, 2009, pp. 305-308.

[10] D. Hanyurwimfura, Y. Liu, and Z. Liu, "Text format based relational database watermarking for non-numeric data," in *proc. IEEE ICCDA,2010*, pp. 312-316.

[11] H. Cui, X. Cui, and M. Meng, "A public key cryptography based algorithm for watermarking relational databases," in *IEEE proc. Of International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2008, pp. 1344-1347.

[12] B. Schneier, *Applied Cryptography, protocols, algorithms and source* code *in C*, 2 nd ed. Wiley-India, 2008, ch. 16, pp. 369-395.

[13] R. Sion, S. M. Atallah, and S. Prabhakar, "Rights protection for relational data," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1509-1525, 2004.

Vidhi Khanduja received her B.Tech degree in Information Technology from Guru Govind Singh Indraprashtha University, Delhi, India and M.E degree in Computer Technology and Applications from Delhi College of Engineering under University of Delhi, Delhi, India and is currently persuing her PhD from Netaji Subhas Institute of Technology, University of Delhi, Delhi, India.

She was Assistant Professor in Department of Information Tcehnology, Delhi College of Engineering (now Delhi Technological University) Delhi, India fro more than 4 years. She is currently working as TRF at Netaji Subhas Institute of Technology, University of Delhi, Delhi, India. She is a member of IACSIT.

Om Prakash Verma received his B.E. degree in Electronics and Communication Engineering from Malaviya National Institute of Technology, Jaipur, India in 1991 and M. Tech. degree in Communication and Radar Engineering from Indian Institute of Technology (IIT), Delhi, India, in 1996 and PhD(S) from University of Delhi, Delhi, India in 2011.

From 1992 to 1998 he was assistant professor in Department of Electronics & Communication Engineering, at Malaviya National Institute of Technology, Jaipur, India. He joined Department of Electronics & Communication Engineering, Delhi College of Engineering (now Delhi Technological University) Delhi, India, as Associate Professor in 1998. Currently, he is Head of Department of Information Technology at Delhi Technological University, Delhi. He is also the author of more than 15 publications in both international journal and conference proceedings. He has guided more than 15 M. Tech. student for their thesis. He has authored a book on Digital Signal Processing in 2003. He is a Principal investigator of an Information Security Education Awareness project, sponsored by Department of Information Technology, Government of India. His research interests include image processing, application of fuzzy logic in image processing, application of evolutionary algorithm in image processing, artificial intelligent and digital signal processing.

Multimed Tools Appl
DOI 10.1007/s11042-012-1037-9

Joint watermarking and encryption for still visual data

**Nidhi Taneja · Gaurav Bhatnagar ·
Balasubramanian Raman · Indra Gupta**

Abstract Joint watermarking and encryption is an upcoming security solution that combines leading but complementary techniques to achieve an enhanced security level. Real time applications using joint watermarking and encryption framework has three requirements: data to be efficiently compressed, watermarking technique to sustain compression, and encryption technique to be developed in a way so as not to disturb the compression efficiency. Finding an optimal solution that combines the three techniques while fulfilling these requirements is a challenging problem. This paper thus, proposes a wavelet domain based joint watermarking and encryption framework that employs singular value decomposition based watermark embedding and sign bit encryption prior to compression. The varying significance of different subbands has been considered to encrypt the data without adversely effecting the compression ratio. Experimental analysis using various evaluation parameters and attack scenarios has revealed the ability of the proposed framework to prove content-ownership, even from the encrypted data. Comparative analysis with the existing techniques reflect its ability to provide better security with less computational

N. Taneja (✉) · I. Gupta
Department of Electrical Engineering, Indian Institute of Technology Roorkee,
Roorkee 247 667, India
e-mail: nidhi.iitr@gmail.com

I. Gupta
e-mail: indrafee@iitr.ernet.in

G. Bhatnagar
Department of Electrical and Computer Engineering, University of Windsor,
Windsor, ON, Canada
e-mail: goravdma@iitr.ernet.in

B. Raman
Department of Mathematics, Indian Institute of Technology Roorkee,
Roorkee 247 667, India
e-mail: balarfma@iitr.ernet.in, balaiitr@ieee.org

Published online: 17 March 2012

resources. This makes it a preferable solution for data security at all stages of data archival, transmission or distribution.

Keywords Joint watermarking and encryption · Singular value decomposition · Set partitioning in hierarchical trees

1 Introduction

With the advancements in the field of communication, coding and networking technology, multimedia applications have increased in day-to-day life. This technological advancement has also equipped potential attackers with the tools to illegally copy, manipulate, or distribute digital data. Hence, security technqiues have become an integral component of data archival, transmission or distribution. This has led to the development of various techniques covered under the umbrella of digital rights management [2, 3, 5–7, 9].

Among the several digital rights management techniques, encryption [13, 16] and watermarking [8, 20] are considered as the first and second line of defence, respectively. The former ensures confidentiality by making the data unintelligible for an unauthorized user, whereas the latter provides copyright protection by embedding a watermark into media data.

Though these two techniques have been developed independently and are complementary to each other, they have been integrated for secure data storage or transmission [10–12, 19, 21]. Their integration not only provides data confidentiality but also proves content ownership at all stages of data consumption. Wu et al. proposed to selectively watermark MPEG data and then encrypt watermarked data [21]. Simitopoulos et al. proposed to embed the watermark in quantized DCT coefficients prior to I-frame encryption using IDEA [19].

To save computational resources, Lian et al. proposed commutative watermarking and encryption technique that perform both these operations in a single step [10]. After identifying varying significance of different parts of image data, middle level wavelet coefficients have been used for watermark embedding, and remaining coefficients (low and high level) for AES encryption. This has also been extended to MPEG data, where residual, MVD and IPM frames are selected for watermarking and encryption, respectively [11]. However, commutative watermarking and encryption is prone to replacement attack due to the mutually exclusive watermarking and encryption data components [12]. A quasi-commutative approach of watermarking and encryption is thus proposed that watermarks and encrypts the entire data to make joint watermarking and encryption (JWE) framework cryptographically secure [12].

Though several frameworks, integrating the two techniques have been developed; it is still in its infancy stage [1, 10–12, 19, 21]. Joint watermarking and encryption (JWE) frameworks are being developed, researched and discarded at a fast pace. Several intricacies are observed in the JWE framework owing to the fact that compression, which is an integral part of encryption, is a potential attack for the embedded watermark.

For a secure multimedia system, an optimal JWE framework requires a clever interweaving of encryption, watermarking and compression. An efficient JWE framework should provide robustness to the embedded watermark against compression without deteriorating the compression efficiency.

The present work, thus, intends to develop a JWE framework that achieves data confidentiality, proves content ownership and offers high compression ratio. In the proposed framework, watermarking and encryption are implemented at content owner and content distributor end, respectively. To achieve the desired objectives, watermark is embedded using singular value decomposition of the wavelet packet transformed image and encryption is performed during SPIHT encoding. Security attained by the proposed JWE framework is ascertained by detailed experimental analysis.

2 Singular value decomposition

Singular value decomposition (SVD) is a powerful technique in many matrix computations and analyses [4]. Use of SVD in matrix computations provides robustness provides robustness against numerical errors. SVD of a square or a rectangular matrix of size $M \times N$ can be expressed as

$$A = U * S * V^T \tag{1}$$

where U and V are orthogonal (unitary) matrices, and S is a diagonal matrix given by $S = diag(\sigma_1, \sigma_2, ..., \sigma_r)$. Here, σ_i denotes singular value of matrix A, and $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$, $1 \leq i \leq r$ and $r = \min(M, N)$. The first r columns of V and U are termed as right and left singular vectors, respectively.

The main motivation for using the SVD is its energy compaction property and its ability to adapt to the variations in local statistics of an image. Each singular value of the image matrix specifies luminance of the image layer, while corresponding pair of singular vectors specify geometry of the image layer. Therefore slight variations of singular values does not affect visual perception of the cover image.

Also, storing the approximation of a matrix using SVD often results in a significant savings over storing the whole matrix. Singular values of a matrix possess algebraic and geometric invariance to some extent, due to which it has certain distinct advantages in digital image processing. For instance, singular values of an image matrix remain same, irrespective of the transposition, rotation or translation performed on the original matrix. Further, singular values of an image are less effected in case of general image processing operations on the image matrix.

3 Proposed joint watermarking and encryption framework

In the proposed JWE framework, the original image X is initially watermarked using key K_w. This watermarked image is then partially encrypted with key K_e. The final image obtained by implementing these two processes in a sequential manner is mathematically expressed as

$$Y = W(X, B, K_w) \tag{2}$$

$$Z = E(Y, K_e) \tag{3}$$

Here, Y, B, K_w, $W()$, Z, K_e and $E()$ are the watermarked copy of original image X, watermark, watermark key, watermark embedding algorithm, encrypted copy of watermarked component Y, encryption key and encryption algorithm, respectively.

The watermarking key, K_w comprises of watermark strength, α^θ as the main key component. This controls perceptibility of the embedded watermark; higher the value of α^θ, more observable is the watermark. An optimal value can be chosen as per the desired visibility of embedded watermark. In contrast, the encryption key, K_e consist of the compression ratio, number of Arnold iterations for scrambling and a seed value for generating a random vector.

A block diagram depicting the proposed JWE framework is indicated in Fig. 1 and the two processes controlled by the independent keys, K_w and K_e, are explained as follows.

The watermarking process initially transforms the host image into wavelet packet transform (WPT) domain. SVD is then performed on all subbands of the transformed image and the watermark image. For watermark embedding, the obtained singular values are modified using (4)

$$(\sigma_{f_{l,\ p}^{\theta}})^* = \sigma_{f_{l,\ p}^{\theta}} + \alpha^\theta\ \sigma_W \tag{4}$$

where $\sigma_{f_{l,\ p}^{\theta}}$ gives original singular values of the subband, $(\sigma_{f_{l,\ p}^{\theta}})^*$ denotes modified singular values of the subband, σ_W denotes singular values of the watermark image, and α^θ is the watermark strength.

After replacing original singular values by the modified values, inverse SVD is taken. This is followed by inverse WPT to retrieve the watermarked image. This watermarked image is transmitted to the content distributor end. Thereafter, encryption is performed on this watermarked image during SPIHT compression [17].

In SPIHT compression, an image is initially transformed into wavelet domain, and a tree structure is formed. The tree structure is then encoded to obtain a SPIHT compressed bitstream. In the proposed framework, encryption is implemented in wavelet domain, just before the formation of tree structure. Encryption is achieved by scrambling the approximation subband using Arnold cat map [15]. This is followed by sign bit encryption of the scrambled transform coefficients using a stream cipher,

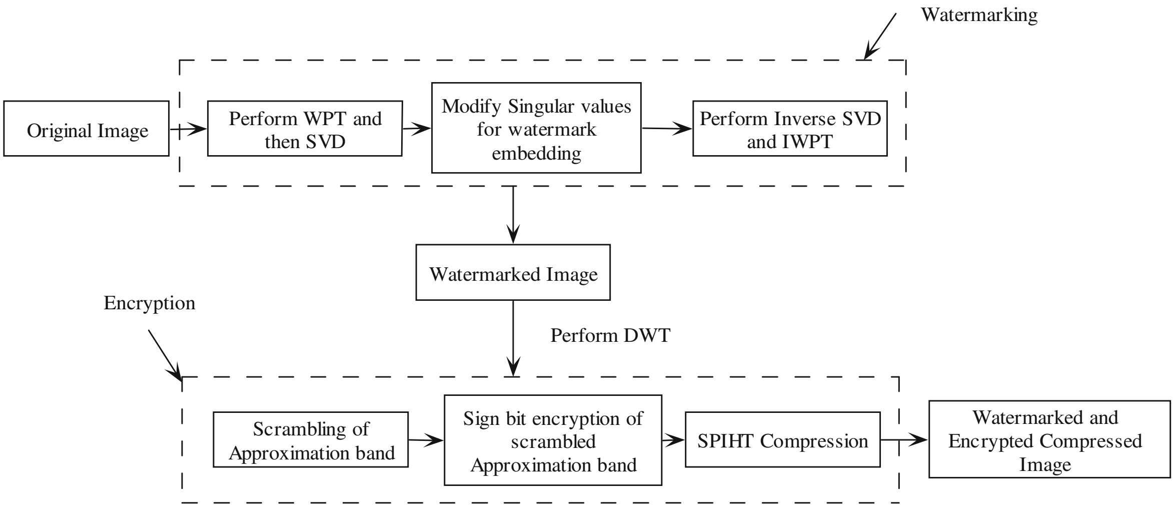

Watermarking
Original Image
Perform WPT and then SVD
Modify Singular values for watermark embedding
Perform Inverse SVD and IWPT
Watermarked Image
Perform DWT
Encryption
Scrambling of Approximation band
Sign bit encryption of scrambled Approximation band
SPIHT Compression
Watermarked and Encrypted Compressed Image

Fig. 1 Block diagram for the proposed framework

generated from a seed value. The original approximation subband is replaced with the scrambled and sign bit encrypted subband. Afterwards, the modified transformed image is used to generate the compressed bitstream using SPIHT.

At the receiver end, compressed bitstream generates the transformed image and inverse DWT of this transformed image provides the reconstructed image. An unauthorized receiver, not having the security keys, would only retrieve an incomprehensible image. Contrary to this, an authorized receiver would perform sign bit decryption and Arnold descrambling of the approximation subband before IDWT. This provides a correctly decrypted output to an authorized receiver. To verify achieved security level of the proposed framework, several experiments have been performed, and are discussed in the next section.

4 Results and discussion

The proposed framework provides twin layer of protection to digital images by combining watermarking and encryption. To substantiate performance of the proposed framework, different subjective and objective evaluation parameters are used. Diverse watermarking and encryption related security attacks are also launched to assess robustness of the proposed framework. Simulations have been performed on various grayscale images, however, results for only 'Barbara' image are illustrated here.

4.1 Subjective and objective evaluation

To examine the quantum of detail actually lost, or retained by the proposed JWE framework, a visual inspection of the watermarked and encrypted images is performed. These images are illustrated in Fig. 2. It is observed that the watermarked

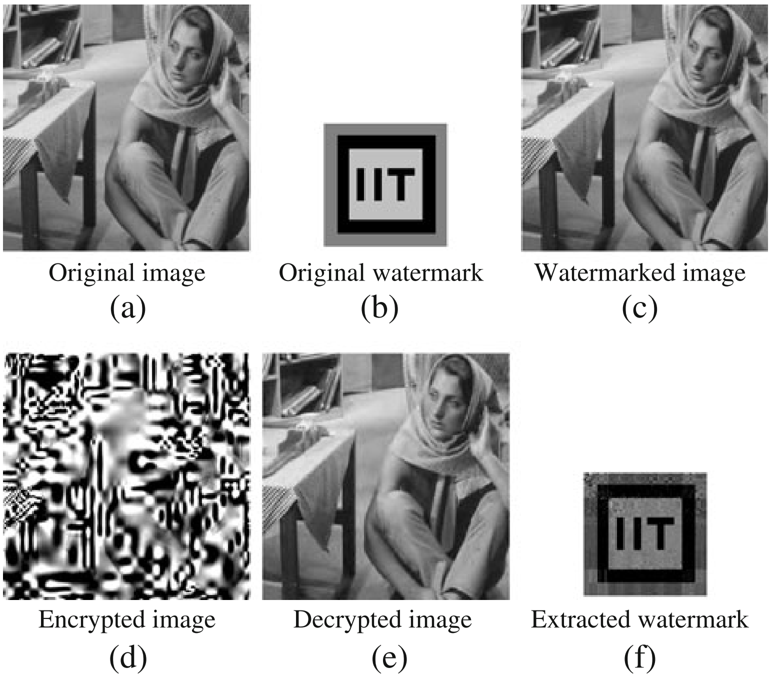


Fig. 2 Results for the proposed framework

Table 1 PSNR (dB) obtained for various images

Image	Barbara	Lena	Plane	Crowd	Bridge	Lake
Watermarked	40.1201	38.7121	38.9503	39.5840	37.5025	37.2499
Encrypted	3.2341	3.1203	3.1842	3.0001	3.1981	2.9064

image is similar to the original image, and the encrypted image is completely incomprehensible. The embedding of watermark in an imperceptible manner has not resulted into loss of any detail from the original image. In contrast, an unintelligible encrypted image reflects that the developed encryption technique provides high data confidentiality and does not leak any information of the original image. To further verify the results, objective evaluation is performed using peak signal to noise ratio (PSNR).

The obtained PSNR values for the watermarked and the encrypted output with reference to the original image are indicated in Table 1. A high PSNR value after watermarking indicates perceptual similarity between the original and the watermarked image, while a low PSNR value of the encrypted output indicates sufficient dissimilarity between the original and the encrypted image. This indicates more computational effort required by an intruder to retrieve the correct image without the knowledge of security keys. PSNR values and visual inspection of results depicts that the proposed technique satisfies the subjective and objective evaluation metric for an acceptable watermarking and encryption technique.

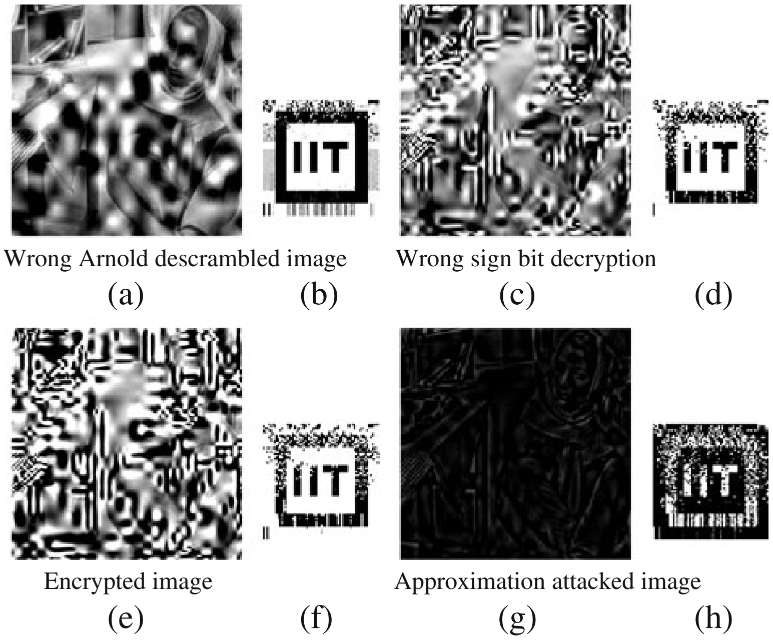
4.2 Key sensitivity analysis

As per the Kerckhoff's principle, security keys are the most important part of any cryptosystem, and decryption using an incorrect key or an approximately correct key should not reveal any details of the original image [18]. To determine key sensitivity of the proposed encryption technique, wrong decryption keys are generated by introducing slight modifications in Arnold scrambling iterations and the seed value. Decryption is then performed by using these slightly modified keys. Figure 3a and c demonstrates the decrypted results when incorrect descrambling iterations or incorrect seed value is considered. It is observed that the images decrypted with wrong decryption keys do not give a clear view of the original image. This reflects high key sensitivity of the developed encryption algorithm.

Thereafter, watermark is extracted from these incorrectly decrypted images. The extracted watermarks are shown in Fig. 3b and d. It is observed that despite unintelligible decrypted images, meaningful watermarks are extracted. This reflects robustness of the watermark embedding technique. To further verify strength of embedding technique, watermark is extracted from the encrypted image. The extracted watermark is indicated in Fig. 3f, and can easily be related to the original watermark.

Strength of the encryption technique is evident from achieved data confidentiality and high key sensitivity. In addition to this, extraction of watermark from the encrypted and incorrectly decrypted image illustrates strength of the watermarking technique. This depicts that content ownership can be proved in a scenario, where a

Fig. 3 Key sensitivity results for the encryption technique


IIT
Wrong Arnold descrambled image
(a)
(b)
Wrong sign bit decryption
(c)
(d)
Encrypted image
(e)
(f)
Approximation attacked image
(g)
(h)

pirate captures an unclear but watermarked copy of the original image. The above analysis corroborates strength of the developed JWE framework.

4.3 Compression performance analysis

In the proposed JWE framework, compression ratio achieved by the employed SPIHT encoder is analyzed. To experimentally evaluate the effect on compression efficiency, original and encrypted images are compressed with 0.8 bits per pixel. As the output bit rate is equal for both the images, length of the compressed bitstream is observed to be same for all the original and the encrypted image. This indicates that the proposed framework does not adversely effect compression efficiency of the SPIHT encoder.

4.4 Approximation attack

Security of the proposed technique is also verified against approximation attack [14]. In this attack, part of the encrypted data is replaced by random data and reconstruction is performed using this partially assumed data. In the present case, few transform coefficients of approximation subband are replaced by a constant value '0', before IDWT. Figure 3g and h shows the reconstructed image and the extracted watermark for this case. It is observed that a clear view of the original image is not obtained. However, watermark extracted from this approximate image has perceptual resemblance with the original watermark. This indicates resistance of the proposed framework for approximation attack.

An approximated copy of the test image is used to measure block-based Luminance Similarity Score (LSS), which captures the coarse luminance information [14]. LSS measures the perception-oriented distance between the clear-text copy of multimedia and attacker's recovered copy from the encrypted media. It was assumed that two given images are pre-processed to be aligned and scaled to the same size. These two images are first divided into blocks in the same way, using 8 × 8 or 16 × 16 non-overlapping blocks. Average luminance values of ith block is then calculated from both images to measure LSS using

$$LSS \cong \frac{1}{N}\sum_{i=1}^{N} f(x_{1i}, x_{2i}) \tag{5}$$

Here, the function $f(x_1, x_2)$ for each pair of average luminance values is defined as

$$f(x_1, x_2) = \begin{cases} 1, & \text{if } |x_1 - x_2| < \frac{\beta}{2} \\ -\alpha\ round\left(\frac{|x_1-x_2|}{\beta}\right), & \text{otherwise} \end{cases} \tag{6}$$

where the parameters α and β control sensitivity of LSS and set to 0.1 and 3, respectively. For the proposed framework, negative LSS is obtained. This indicates a substantial dissimilarity in luminance of the two images.

4.5 Attack analysis

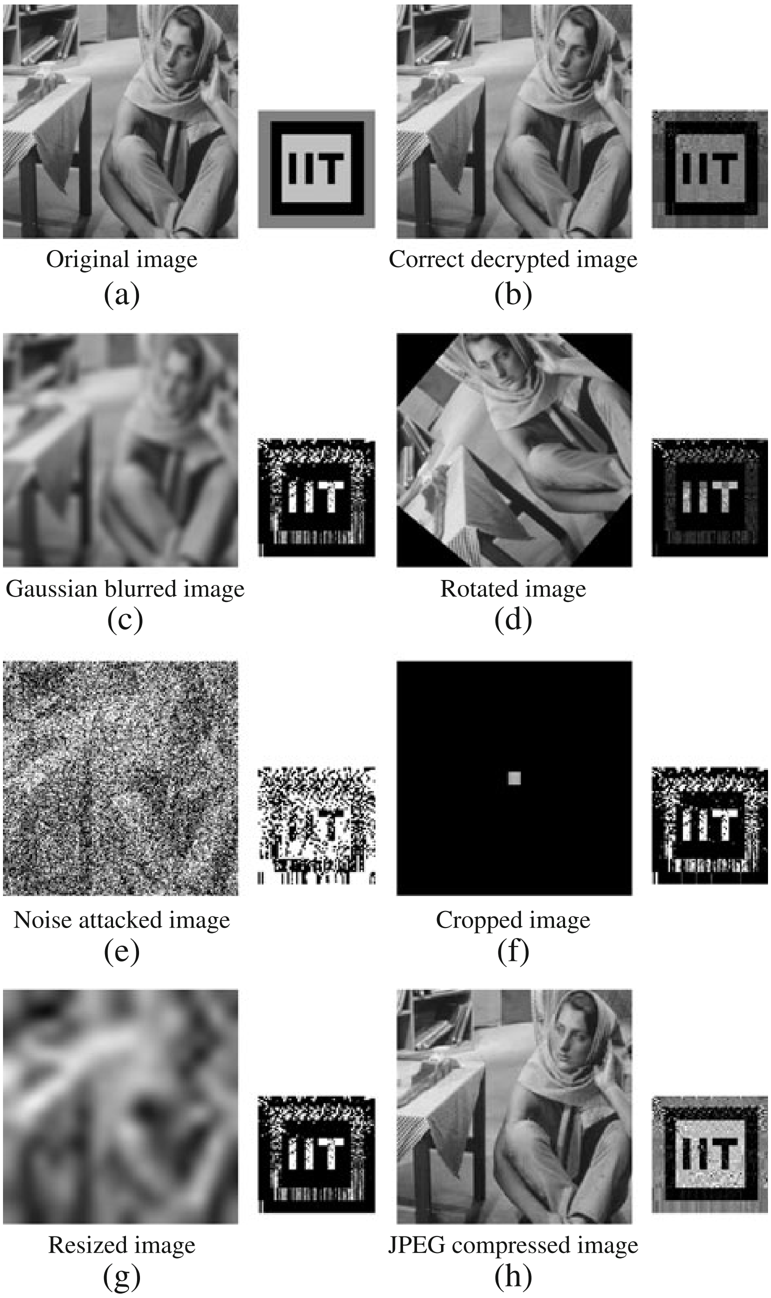
After performing key sensitivity and approximation attack analysis for the proposed encryption technique, this section discusses performance of the proposed watermarking technique in different attack scenarios. Robustness of the proposed watermarking technique is investigated by launching various attacks on the watermarked image, and observing the quality of extracted watermarks from the attacked images. Visual inspection of the extracted watermark is performed to assess its perceptual similarity with the original watermark. For an objective evaluation of similarity, correlation coefficient is calculated between the extracted and actual singular values, using

$$\rho(w, \bar{w}) = \frac{\sum\limits_{i=1}^{r} \left(w(i) - w_{\text{mean}}\right)\left(\overline{w}(i) - \overline{w}_{\text{mean}}\right)}{\sqrt{\sum\limits_{i=1}^{r} \left(w(i) - w_{\text{mean}}\right)^2}\sqrt{\sum\limits_{i=1}^{r} \left(\overline{w}(i) - \overline{w}_{\text{mean}}\right)^2}} \tag{7}$$

where w, $\overline{w}$, w_{mean} and $\overline{w}_{\text{mean}}$ are the original singular values, extracted singular values, mean of the original singular values and mean of the extracted singular values. Here, $r = \min(M, N)$, and (M, N) denote size of the image.

Among the various attacks launched on the watermarked image, basic attack includes (a) a 13 × 13 Gaussian blurring on the watermarked image, (b) rotation of the watermarked image by 50°, and (c) addition of 80% Gaussian noise to the watermarked image. Watermarks are extracted from these three attacked images. Figure 4c–e indicates the attacked watermarked images and their corresponding extracted watermarks. It is observed that the extracted watermarks are recognizable and can be assumed as a degraded version of the original watermark. Correlation coefficient values for the extracted watermarks is indicated in Table 2.

As cropping is a frequently used operation in image applications, watermarked image is also tested for cropping attack. Process of selecting and removing a portion of an image is generally performed to create focus or strengthen the composition. In the present test case, the watermarked image is cropped to only 2.5% of the actual



Original image
(a)

Correct decrypted image
(b)

Gaussian blurred image
(c)

Rotated image
(d)

Noise attacked image
(e)

Cropped image
(f)

Resized image
(g)

JPEG compressed image
(h)

Fig. 4 Image and its extracted watermark

size, and watermark extraction is performed. Figure 4f indicates the cropped image and the extracted watermark. It is to be noted that the watermark could be extracted, even from an image equal to 2.5% of the actual image size.

Another frequently used image processing operation is resizing, wherein the image is reduced or enlarged to a desired size. This leads to data loss of the original

Table 2 Correlation coefficient (CC) for extracted watermark from attacked Barbara image

Attack	Gaussian Blur	Rotation	Noise	JPEG compression	Resizing	Cropping
CC	−0.6885	−0.9402	0.3732	0.9656	−0.6832	−0.6079

image and the watermark embedded within it. In the present test case, the image is reduced to 16 × 16 and again carried back to the original size 256 × 256. Figure 4g depicts the resized image and its corresponding extracted watermark. It is observed that the extracted watermark is still recognizable and is similar to the original watermark.

Another potential attack for a watermarking technique is compression, that is generally performed owing to the large data size and limited channel bandwidth. As image compression techniques are lossy in nature, they lead to data loss from the entire image and the watermark embedded in it. Despite the losses, a secure system requires that the watermark is extractable, even from a compressed image.

To assess the proposed technique against compression attack, lossy JPEG compression, with a compression ratio of 80:1, is performed on the watermarked image. Watermark is extracted from this JPEG compressed image. Figure 4h illustrates the JPEG compressed image and its extracted watermark. It is observed that the extracted watermark is of very high quality, and almost an exact replica of the original watermark. Further, it is to be noted that the proposed framework is based on SPIHT compression of the watermarked data. Hence, the watermark indicated in Fig. 4b is actually the watermark extracted from a SPIHT compressed image. This demonstrates robustness of the proposed watermarking technique against SPIHT compression. This reflects that the proposed watermarking technique can withstand lossy JPEG and SPIHT compression attack.

In all the above-mentioned attack scenarios, it is observed that the extracted and the original watermark are perceptually similar. This, along with the correlation coefficient values indicated in Table 2, reflects the resistance of the proposed watermarking technique against various image processing operations. The above-discussed analysis reflects the ability of the proposed framework to prove content ownership, even from attacked and compressed images.

5 Comparative analysis

This section discusses comparative analysis of the proposed framework, with an existing JWE framework [10]. The existing framework utilizes the entire image data to provide data confidentiality. It provides AES encryption to all the coefficients of low level subband, and sign bit encryption to all the coefficients of remaining subbands. In contrast, the proposed framework performs only sign bit encryption

Table 3 Scheme of existing and proposed JWE

Technique	Watermarking	Encryption
Existing [10]	All coefficients of middle level subband are used, i.e., $(M_1 \times N_1)$ coefficients used for a subband of size $(M_1 \times N_1)$	AES in approximation subband and sign bit encryption in remaining subbands
Proposed	Uses only singular values for watermarking i.e., min. (M_1, N_1) coefficients used for a subband of size $(M_1 \times N_1)$	Scrambling and sign bit encryption of only approximation subband

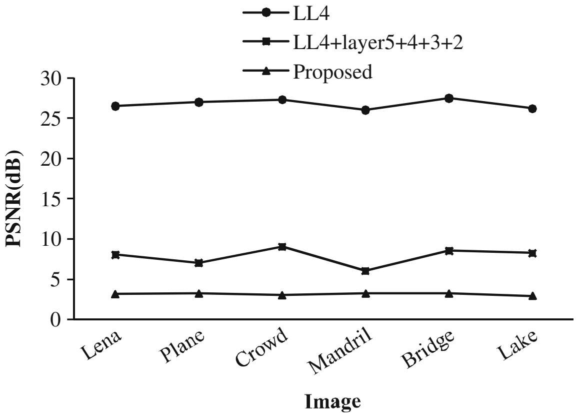


Fig. 5 PSNR comparison chart

for coefficients of the approximation band. This drastically reduces the amount of data encrypted in the proposed technique.

Further, existing technique uses all the coefficients of middle level subband for watermarking [10]. The proposed framework, however, performs watermark embedding, only in the singular values of the subband. As singular values form a diagonal matrix, thus, for a subband of size $(M_1 \times N_1)$, the coefficients used for watermarking with the proposed technique are expressed as $\min.(M_1, N_1)$. This reflects a drastic reduction in computational requirements, as compared to $(M_1 \times N_1)$ coefficients, that are used for the existing technique [10].

A comparative representation for the amount of data used for watermarking and encryption is indicated in Table 3. Apart from the amount of data used, existing and proposed techniques are also compared on the basis of attained PSNR value. A comparative graph for PSNR value of the proposed and existing techniques is depicted in Fig. 5. This reveals low PSNR value achieved by the proposed framework, which is analogous to high data confidentiality.

To summarize, comparative analysis indicate that despite the small quantum of data watermarked and encrypted in the proposed framework, PSNR obtained is better than the existing techniques.

6 Conclusion

JWE is emerging as an effective security solution, that provides data confidentiality and proves content ownership. Compression is necessary for encryption; however, it behaves as an attack for the embedded watermark. In such a scenario, obtaining an optimal solution is a dexterous task.

A novel JWE framework is thus presented that performs watermarking and encryption in an independent manner at the content owner and distributor end, respectively. In the proposed framework, watermark is embedded in the wavelet packet domain using SVD, and the watermarked image is partially encrypted during SPIHT compression. Thorough performance analysis reflects the robustness of the proposed framework to withstand compression, approximation and various other image processing attacks. The developed framework does not adversely effect the compressibility of the SPIHT encoder. The ability to prove content ownership, even from a compressed, encrypted, or an attacked image is also validated. Thus, it acts as a two-fold impediment to illegal distribution of media data, and a preferable choice for secure image transmission or distribution.

References

1. Boato G, Conci N, Conotter V, De Natale FGB, Fontanari C (2008) Multimedia asymmetric watermarking and encryption. Electronics Lett 44(9):601–602
2. Chang FC, Huang HC, Hang HM (2007) Layered access control schemes on watermarked scalable media. J VLSI Signal Process Syst Signal Image Video Technol 49(3):443–455
3. Committee on Intellectual Property Rights in the Emerging Information Infrastructure (2000) The digital dilemma: intellectual property in the information age. US National Research Council, National Academic Press, Washington, D.C.
4. Dewilde P, Deprettere EdF (1988) Singular value decomposition. An introduction. In: Deprettere EdF (ed) SVD and signal process.: algorithms, applications, and architectures. Elsevier Science Publishers, North Holland, pp 3–41
5. Eskicioglu AM (2003) Protecting intellectual property in digital multimedia networks. IEEE Comput 36(7):39–45
6. Eskicioglu AM (2003) Multimedia security in group communications: recent progress in key management, authentication, and watermarking. Multimedia Syst 9:239–248
7. Huang HC, Chen YH (2009) Genetic fingerprinting for copyright protection of multicast media. Soft Comput 13(4):383–391
8. Kundur D, Hatzinakos D (2004) Towards robust logo watermarking using multiresolution image fusion. IEEE Trans Multimedia 6(1):185–197
9. Li B, He J, Huang JW, Shi YQ (2011) A survey on image steganography and steganalysis. J Inf Hiding Multimedia Sig Proc 2(2):142–172
10. Lian S, Liu Z, Zhen R, Wang H (2006) Commutative watermarking and encryption for media data. Optical Engg Lett 45(8):1–3
11. Lian S, Liu Z, Ren Z, Wang H (2007) Commutative encryption and watermarking in video compression. IEEE Trans Circuits Syst Video Technol 17(6):774–778
12. Lian S (2009) Quasi-commutative watermarking and encryption for secure media content distribution. Multimed Tools Appl 43(1):91–107
13. Liu J-L (2006) Effective selective encryption for Jpeg2000 images using private initial table. Pattern Recogn 39:1509–1517
14. Mao Y, Wu M (2006) A joint signal processing and cryptographic approach to multimedia encryption. IEEE Trans Image Process 15(7):2061–2075
15. Peterson G (1997) Arnold's cat map. Available from: http:online.redwoods.cc.ca.us/instruct/darnold/maw/catmap3.htm
16. Pommer A, Uhl A (2003) Selective encryption of wavelet-packet encoded image data: Efficiency and security. Multimedia Syst 9(3):279–287
17. Said A, Pearlman WA (1996) A new, fast, and efficient image codec based on set partitioning in hierarchical trees. IEEE Trans Circuits Syst Video Technol 6(3):243–250
18. Schneier B (1995) Applied cryptography second edition: protocols, algorithms, and source code in C. Wiley, New York
19. Simitopoulos D, Zissis N, Georgiadis P, Emmanouilidis V, Strintzis MG (2003) Encryption and watermarking for the secure distribution of copyrighted MPEG video on DVD. Multimedia Security 9(3):217–227
20. Su K, Kundur D, Hatzinakos D (2005) Statistical invisibility in collusion-resistant digital video watermarking. IEEE Trans Multimedia 7(1):43–51
21. Wu T, Wu S (1997) Selective encryption and watermarking of MPEG video. In: Proc. int. conf. image science, systems and technology. Ontario, Canada

Nidhi Taneja received her B.E. Degree in Electronics & Communication and M.Tech Degree in Digital Communication in 2001 and 2006, respectively. She has received her Ph.D. in Electrical Engineering at Indian Institute of Technology Roorkee, India. At present, she is an Assistant Professor in Department of Electronics and Communication Engineering at Delhi Technological University (Formerly Delhi College of Engineering), New Delhi. Her area of interest includes Wireless Communication, Multimedia Transmission over Packet Networks, Image Encryption, Watermarking, Biometrics and Visual Cryptography.

Gaurav Bhatnagar received his Ph.D. in Mathematics from Indian Institute of Technology Roorkee. At present, he is a post-doctoral fellow in Department of Electrical and Computer Engineering at University of Windsor, Canada. He has several research papers in various reputed international journal and conferences. His areas of research include Image Analysis, Image Fusion, Biometrics, Wavelet Analysis, Cryptography and Digital Watermarking.

Balasubramanian Raman received his Ph.D. in Mathematics (2001) from Indian Institute of Technology, Madras, India. At Present, he is an Assistant Professor and Head of the Computer Vision, Graphics and Image Processing Laboratory in the Department of Mathematics at Indian Institute of Technology Roorkee, India. He worked as a Post Doctoral Associate in ECE Department, and member of the Visualization Research Laboratory (VIZ Lab), at Rutgers, The New State University. He was also a Post Doctoral fellow of Computer Engineering and Computer Science (CECS), and member of the Computational Intelligence Research Laboratory (CIRL), at the University of Missouri-Columbia (MU), Missouri, USA. He has also worked as Visiting Professor in Department of Electrical and Computer Engineering at University of Windsor, Canada under Boyscast Fellowship. His area of research includes Computer Vision, Graphics, Satellite Image Analysis, Scientific Visualization, Imaging Geometry, Reconstruction Problems, Image Encryption and Digital Watermarking.

Indra Gupta received her B.Tech. Degree in Electrical Engineering from HBTI, Kanpur, in 1984. She completed her M.E. and Ph.D. from University of Roorkee, India. Currently, she is an Associate Professor in the Department of Electrical Engineering, Indian Institute of Technology Roorkee, India. Her areas of interest includes Advanced Microprocessor Applications, Information Security, Multimedia Processing, Process Control Applications, Biomedical Imaging, Content based Image Retrieval and Online Computer Applications.

PERFORMANCE AND EXHAUST GAS EMISSIONS ANALYSIS OF DIRECT INJECTION CNG-DIESEL DUAL FUEL ENGINE

RANBIR SINGH[1*]

Research Scholar, PhD Candidate

University of Delhi, Delhi, INDIA

ranbirsharma2812@gmail.com

SAGAR MAJI[2]

Professor, Research Guide

Delhi Technological University, Bawana Road, Delhi - 110042, INDIA

smaji321@yahoo.com

[1*] Corresponding author, email: ranbirsharma2812@gmail.com, Contact No. : 09868271179

Abstract:

Existing diesel engines are under stringent emission regulation particularly of smoke and particulate matter in their exhaust. Compressed Natural Gas and Diesel dual fuel operation is regarded as one of the best ways to control emissions from diesel engines and simultaneously saving petroleum based diesel fuel. Dual fuel engine is a conventional diesel engine which burn either gaseous fuel or diesel or both at the same time. In the present paper an experimental research was carried out on a laboratory single cylinder, four-stroke variable compression ratio, direct injection diesel engine converted to CNG-Diesel dual fuel mode to analyze the performance and emission characteristics of pure diesel first and then CNG-Diesel dual fuel mode. The measurements were recorded for the compression ratio of 15 and 17.5 at CNG substitution rates of 30% and 60% and varying the load from idle to rated load of 3.5kW in steps of 1 up to 3kW and then to 3.5kW. The results reveal that brake thermal efficiency of dual fuel engine is in the range of 30%-40% at the rated load of 3.5 kW which is 11%-13% higher than pure diesel engine for 30% and 60% CNG substitution rates. This trend is observed irrespective of the compression ratio of the engine. Brake specific fuel consumption of dual fuel engine is found better than pure diesel engine at all engine loads and for both CNG substitution rates. It is found that there is drastic reduction in CO, CO_2, HC, NO_x and smoke emissions in the exhaust of dual fuel engine at all loads and for 30% and 60% CNG substitution rates by employing some optimum operating conditions set forth for experimental investigations in this study.

Key Words: Compressed Natural Gas; Dual Fuel Engine; Emission Analysis; Variable Compression Ratio Engine; Brake Thermal Efficiency; Brake Specific Fuel Consumption.

1. INTRODUCTION

Impending possible energy crisis in future, rising costs and toxic emissions associated with conventional petroleum fuels have caused researchers to search out and investigate the possibility of utilization of alternate clean and non-polluting gaseous fuels for internal combustion engines. Existing diesel engines are under stringent emission regulation particularly of smoke and NO_X in their exhaust. Much interest has centered on

Compressed Natural Gas due to its potential for low particulate and NO_X emissions. Compressed Natural Gas and diesel dual-fuel operation is regarded as one of the best ways to control emissions from diesel engines and simultaneously save petroleum based precious diesel fuel. Dual fuel engine is a conventional diesel engine which burn either gaseous fuel or diesel or both at the same time. The mode of operation is defined as straight diesel if only diesel fuel is used and dual fuel if two fuels are used at the same time. In dual fuel operation the gaseous fuel is mixed with air at lean gas-air ratios and the mixture is then compressed during the compression stroke. Near the end of compression stroke, diesel fuel is injected. After a short ignition delay the combustion of diesel occurs first, igniting the natural gas and the flame propagation begins. The introduction of CNG along with intake air changes the thermodynamic and chemical properties of the mixture in the cylinder and thus the dual fuel combustion has its own characteristics on performance and emission characteristics of a dual fuel engine. The diesel fuel which acts as a source of ignition is often referred to as pilot diesel. The quantity of pilot diesel and concentration of CNG in the intake air have important effects on the performance and emissions of a dual fuel engine.

There have been several fundamental studies on dual fuel engines. [Karim G A (1983)] reviewed the prospects, problems and solutions of the dual engine of the CI engine type. [Roydon *et al.* (1991)] studied auto-ignition of pure methane and natural gas in a simulated diesel environment using a constant volume combustion vessel for the pressure and temperature ranges of 5 to 55 atm and 600 to 1700K. [Karim G A (1991)] examined some measures for improving the performance of gas fuelled diesel dual fuel engines at light load. [G.E. Doughty *et al.* (1982)] studied natural gas fueling of a diesel engine and found that for full load operation, fuel efficiency was similar to diesel operation. [Liu Z and Karim G A (1997)] developed simulation model of combustion process in gas-fuelled diesel engines. [Guowai Li *et al.* (1991)] carried out an optimization study of pilot-ignited natural gas direct-injection in diesel engine. [Youtong Z *et al.* (2003)] formulated dual fuel engine simulation model and studied the combustion process of a diesel-natural gas dual fuel engine and good levels of agreement were obtained between measured and predicted results. [Singh S *et al* (2004)] studied the combustion and emissions of a diesel-natural gas dual fuel engine and shown that dual fuel engine combustion results in significant reduction in NO_X and smoke emissions. [Karim G A *et al.* (1980)] and [Xianhua D *et al.* (1986)] have reported that at light load, dual fuel engines usually exhibit a drop in brake thermal efficiency and power output in comparison to pure diesel operation. The emissions of unburned hydrocarbons and carbon monoxide are found higher than neat diesel operation at light loads.

The main objective of the present study is to investigate the effect of Exhaust Gas Recirculation, intake air temperature, rate of injection of pilot fuel quantity, intake air throttling and substitution of CNG at two compression ratios of 15 and 17.5, on the performance parameters and emissions of a CNG-Diesel dual fuel engine.

2. MATERIAL AND METHODS

2.1 CNG as an alternate fuel for internal combustion engines

CNG has emerged as a promising alternative fuel due to its clean burning characteristics and very low amount of exhaust emissions. In petrol engines CNG is used by installing a Bi-Fuel Conversion kit and the converted engine has the flexibility of operation either on CNG or petrol. Diesel engines can also be converted to run on CNG by installing a dual fuel conversion kit or converting the existing diesel engine into SI engine. Most existing CNG vehicles use petrol engines, modified by after-market retrofit conversions and retain bi-fuel capability. Such bi-fueled converted engines generally suffer from a power loss and can encounter drivability problems, due to the design and installation of the retrofit conversion kit. Whereas single fuel engines optimized for CNG are likely to be considerably more attractive in terms of performance and emissions. In diesel engines CNG as a fuel can be used in dual fuel mode and offers the advantage of reduced emissions of NO_X, particulate matter and CO_2 while retaining the thermal efficiency of the conventional diesel engine, [Bhandari K *et al.* (2005)].

The safety aspects of converting engines to run on CNG are of great concern to users of CNG vehicles. However, CNG has four big safety features that make it an inherently safer fuel than petrol, diesel, or LPG. Its specific gravity is 0.587 which means that it is lighter than air and even if it leaks out it just rises up and dissipates into the atmosphere. Its self ignition temperature is 540^0C compared to 227-500^0C for petrol and 257^0C for diesel fuel and higher flammability limits give the gas a high dispersal rate and make the likelihood of fire in the event of a gas leak much less than for petrol or diesel. CNG has to mix with air within small range of 4 to 14% by volume for combustion to occur which is far narrower range than for petrol or diesel fuels. CNG cylinders are designed and built with special materials to the highest safety specifications, which makes its storage far safer than petrol or diesel fuel tanks.

The life of engine increases by using CNG. Lubricating oil life is extended considerably because CNG does not contaminate and dilute the crankcase oil. A big advantage of CNG is that it is virtually pollution free. CNG has a good mixture quality with air and when correct proportions are brought together they mix thoroughly and rapidly, which improves combustion efficiency of the engine. The higher Research Octane Number (130) for CNG as compared to that of petrol (87) allows a higher compression ratio (15.6:1) and consequently more efficient fuel consumption. Due to higher compression ratio Diesel engines can also use CNG as a fuel. But it cannot replace diesel completely like petrol due to poor cetane rating of CNG, [Singh R *et al.* (2012)]. Hence CNG seems a very attractive option for its use in diesel engines. The properties of CNG as a fuel for IC Engines are given in Table 1, [Sera M.A. *et al*. (2003)].

Table1: Properties of CNG as a Fuel at 25^0c and 1 atm.

CNG properties	**Value**
Density (kg/m^3)	0.72
Flammability limits (volume% in air)	4.3-15
Flammability limits (Ø)	0.4-1.6
Auto ignition temperature in air (^{0}C)	723
Minimum ignition energy (MJ)	0.28
Flame velocity (m/sec)	0.38
Adiabatic flame temperature (K)	2214
Quenching distance (mm)	2.1
Stoichiometric fuel/air mass ratio	0.069
Stoichiometric volume fraction (%)	9.48
Lower heating value (MJ/Kg)	45.8
Heat of combustion (MJ/Kg_{air})	2.9

2.2. Development of experimental test set up

A single cylinder, 04 stroke, variable compression ratio, water cooled diesel engine installed at authors internal combustion laboratory was converted to operate on dual fuel mode by carrying out minor modifications. The CNG fuel was mixed with intake air at a point in the intake manifold just outside the cylinder. The test engine is directly coupled to an electric dynamometer, which permits the engine to operate under partial monitoring conditions representing negative brake output. For any set of operating conditions, the pilot fuel was kept constant while the amount of CNG fuel was gradually increased. The ignition delay period was established from records obtained using a water-cooled piezoelectric transducer. The injection timing was established using an electric inductance transducer. The average values obtained from several consecutive cycles were used. During the tests the injection timing was kept constant and the engine was operated at 1000 RPM, under naturally aspirated conditions. In the test set up a number of measuring and other ancillary instruments were used which included: test engine, CNG conversion kit, eddy current type dynamometer, air box with orifice meter and manometer, piezo-sensor range 5000PSI with low noise cable, crank angle sensor, data acquisition device, piezo-powering unit, digital milivoltmeter, temperature sensor, temperature transmitter, load indicator, load sensor, fuel flow transmitter, air flow transmitter, engine performance analysis software, rotameter, exhaust gas analyzer, smoke meter etc. Fig.1 shows a schematic layout of engine test set up and Fig.2 shows actual image of CNG-Diesel dual fuel engine test set up. Test engine specifications are given in Table 2.

Table 2: Specifications of test engine

Engine type	**Make Kirloskar**
Bore	87.5mm
Stroke length	110mm
No. of cylinders	01
No. of strokes	04
Type of cooling	Water cooled
Rated power	3.5 kW at 1500RPM
Engine capacity	661cc
Compression ratio	17.5
Variable CR range	12 to 18
Fuel used	Diesel, CNG-Diesel in dual fuel mode

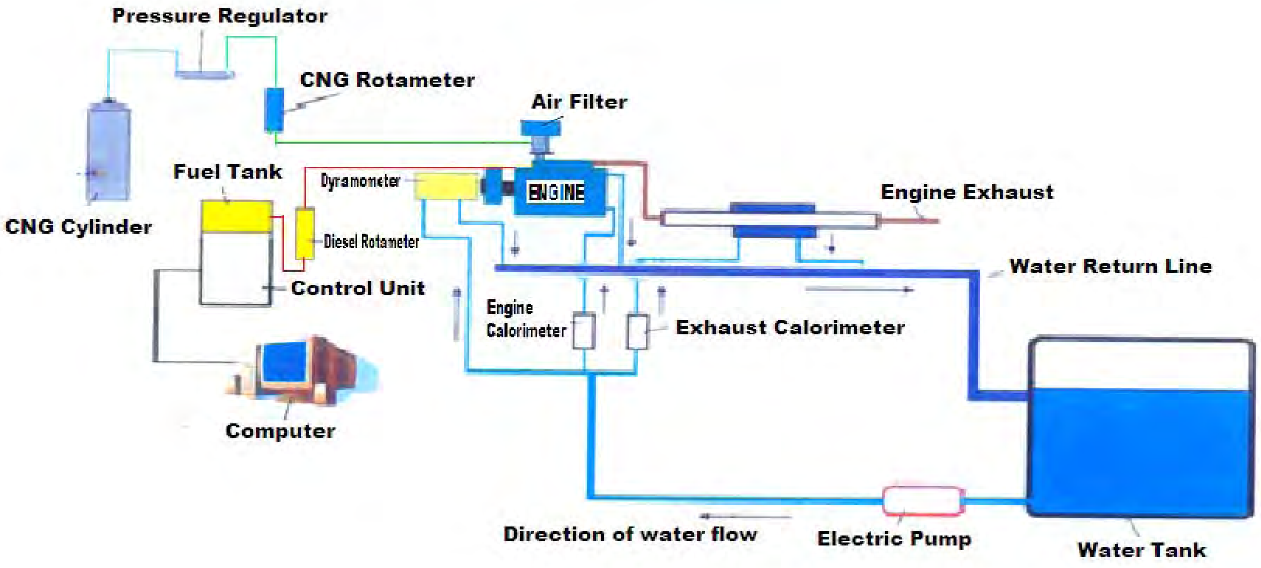


Fig.1: Schematic layout of CNG-diesel dual fuel engine test set up

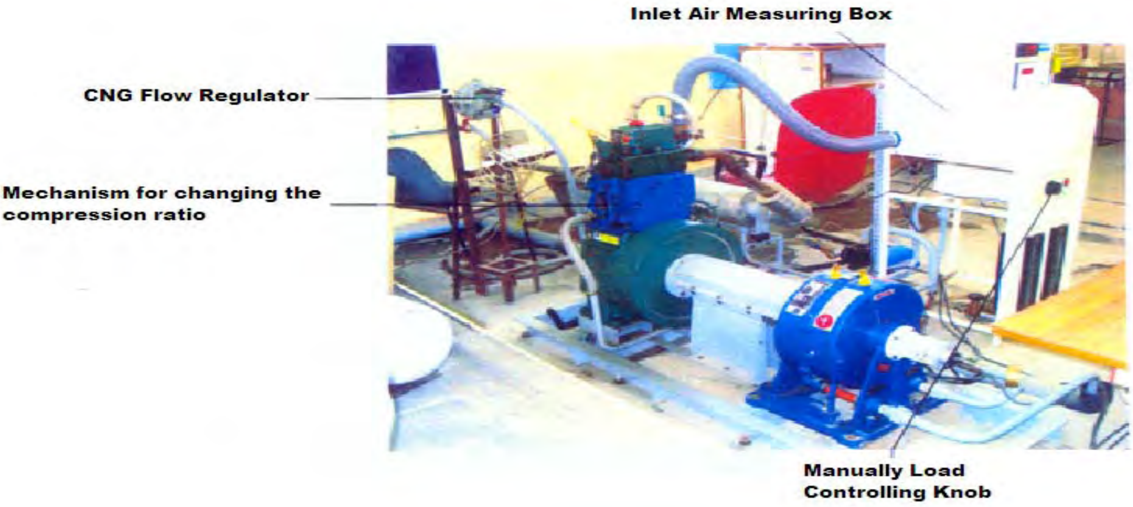

Inlet Air Measuring Box
CNG Flow Regulator
Mechanism for changing the compression ratio
Manually Load Controlling Knob

Fig.2: Actual image of CNG-Diesel dual fuel engine test set up.

2.3. Measurements and observations

A set of reading was obtained first by running the engine with diesel fuel at CR of 17.5 and varying the load from idle to rated load of 3.5kW in steps of 1 up to 3kW and then to 3.5kW. The engine performance parameters were recorded by using the Software engine Soft and other instruments.

The emissions were recorded by using Gas Analyzer (AVL Di Gas 444) and the opacity was recorded by Smoke meter (AVL437). Another set of reading was recorded for the operation of the engine in CNG-Diesel dual fuel mode. For this CNG conversion kit was switched ON and the flow of the CNG substitution rate was set at 30%. Similar set of readings were recorded for the compression ratio of 15 by changing it using the tilting head arrangement. Same set of readings were recorded for CNG substitution rate of 60%. The various performance parameters recorded were: engine load, brake thermal efficiency, brake specific fuel consumption etc. Exhaust gas emissions recorded were: CO in %, CO_2 in%, unburned hydrocarbons (UBHC) in parts per million (PPM), and oxides of nitrogen (NO_X) in PPM by using gas analyzer. Opacity of the smoke in the exhaust was measured in % by using smoke meter. As reported in the literature and during the experiments, it was observed that thermal efficiency of dual fuel engine was lower than pure diesel mode at part and low loads and emissions of CO and HC were observed higher than diesel operation. These were improved by employing larger pilot fuel quantity, using small percentage of EGR (Exhaust Gas Recirculation), increasing intake temperature and adjustment of rate of pilot fuel injection. Table 3, gives the optimum operating conditions set forth for the experimental investigations at different loads for diesel and dual fuel operation modes.

Table3: Optimum operating conditions employed for dual fuel operation.

Load, kW	**Intake Temp. K**	**Pilot fuel quantity, mg/cycle**	**Optimum EGR by volume in %**	**Throttle opening in %**
01	346	9.1	16	40
02	346	09	13	65
03	337	07	07	100
3.5	312	05	03	100

3. RESULTS AND DISCUSSION

The results obtained by performing experiments by employing optimum operating conditions mentioned in table 3, under pure diesel mode and dual fuel mode of operation are compared and analyzed by representing them graphically.

3.1 Brake thermal efficiency analysis

The brake thermal efficiency is plotted against the load applied and the curves for 30% and 60% CNG substitution rates at CR of 15 and 17.5 are plotted together both for pure diesel fuel and dual fuel modes in figures 3 and 4. It can be noticed from figure 3 that value of brake thermal efficiency of dual fuel mode with 30% CNG substitution rate, is more than diesel fuel mode by 5.11%, 5.58%, 9.77% and 10.74% at 1, 2, 3 and 3.5kW loads respectively. For 60%CNG substitution the value of brake thermal efficiency of dual fuel fuel mode is more than neat diesel mode by 9.03%,11.17%,13.43% and 12.6% at 1, 2, 3 and 3.5kW engine loads respectively. Fig.4 depicts the variation of brake thermal efficiency with variation of engine load at CR of 17.5 for the two substitution rates of 30% and 60% and at both these substitution rates, B.TH.E. of dual fuel engine is more than that of pure(0%CNG) diesel operation from no load to full load and this difference is maximum at full load and follow almost the same trend as at CR of 15 with slightly higher value at CR of 17.5.

This increase in the value of brake thermal efficiency of dual fuel engine is low at low loads but significantly high at higher engine loads because at low loads the fuel air ratio of the air-CNG mixture is very low, resulting in incomplete flame propagation and most of the fresh air-gas mixture remains unburnt. At higher engine loads, air-fuel ratio decreases, resulting in complete combustion and increase in brake thermal efficiency.

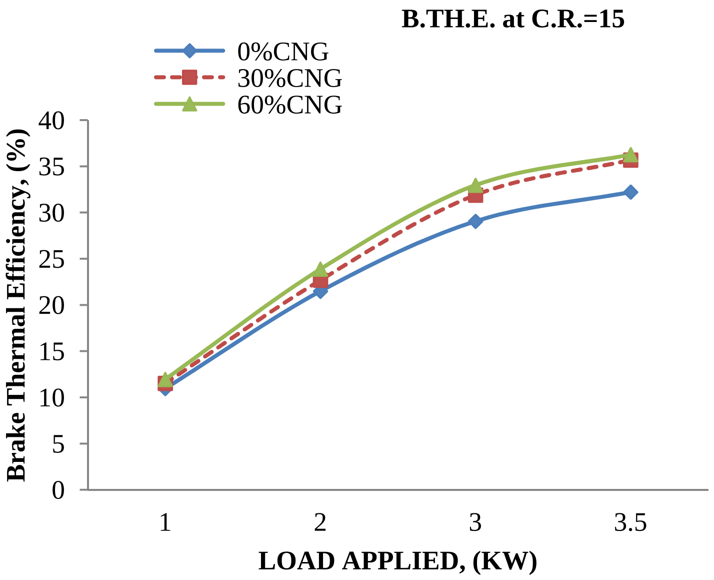


Fig.3: Brake thermal efficiency with engine load for diesel and dual fuel mode at CR=15.

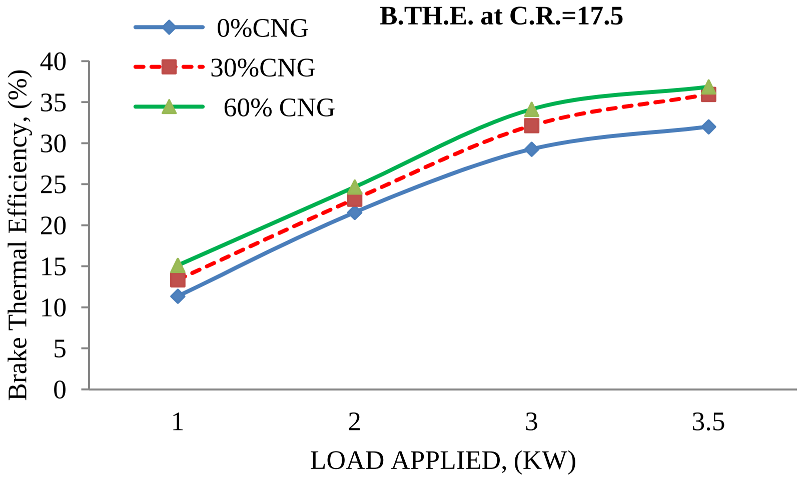


Fig.4: Brake thermal efficiency with engine load for pure diesel and dual fuel mode at C.R. =17.5.

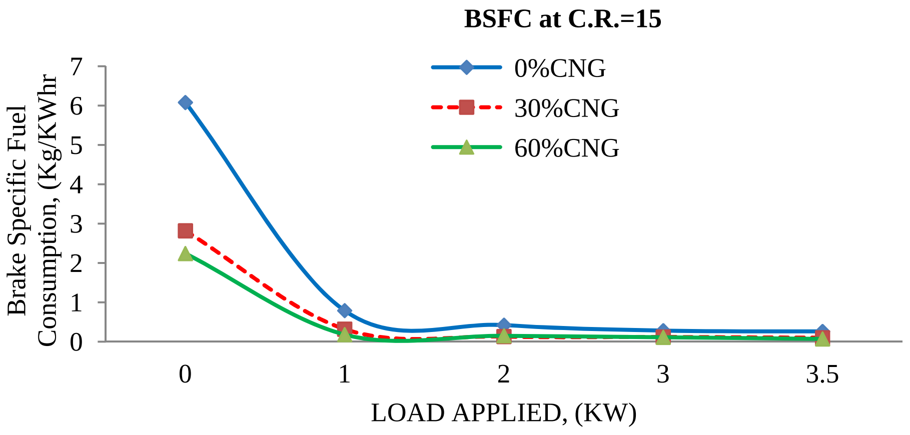


Fig. 5.:Brake specific fuel compustion with load for diesel and dual fuel mode at C.R.=15

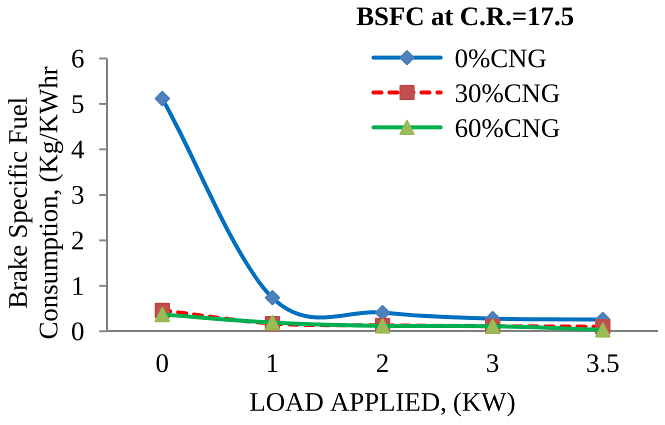


Fig.6: Brake specific fuel combustion with load for pure diesel and dual fuel modes at C.R. =17.5

3.2 Brake specific fuel consumption analysis

It is the consumption of the fuel in kg/kWhr of the brake output of the engine. The observations for BSFC for base diesel and dual fuel modes at CR of 15 and 17.5 and for 30%CNG, 60%CNG substitution rates were recorded and are represented graphically for analysis. From figure 5 it is observed that BSFC with 30%CNG dual fuel mode is less than pure diesel mode by 59.49%, 69.04%, 57.14%, 61.54% at 1, 2, 3 and 3.5kW engine loads respectively at compression ratio of 15. The same trend is observed for 60%CNG substitution rate. As the compression ratio is further increased to 17.5, BSFC value for dual fuel mode is further decreased by 77.02%, 68.29%, 60.71% and 61.54% at 1, 2, 3, 3.5kW loads respectively at 30%CNG substitution rate in comparison to pure diesel mode as depicted in figure6. It further decreases as the substitution rate of CNG is further increased to 60%. With CNG replacing diesel fuel, it contributes to extra heat energy on combustion, resulting in better BSFC than diesel mode. This could also be due to higher calorific value of CNG, better mixing of air and CNG and improved combustion efficiency.

4.EXHAUST GAS EMISSIONS ANALYSIS

Exhaust gas emissions for pure diesel and dual fuel modes were measured experimentally by exhaust gas analyser. The emissions recorded were CO, CO_2, UBHC and NO_X. The smokemeter was used to measure the opacity of the smoke.

4.1 CO Emissions analysis

Figures 7 and 8 show the effect of engine load, CNG substitution rates and changing compression ratio on CO emission concentration in diesel and dual fuel modes. It can be noticed from figure 7, that CO emissions decrease as the load on the engine is increased for diesel and dual fuel modes. CO emissions for 30% and 60% CNG substitution rates are lower than pure diesel mode in the range of 33.3%-61.9% for different engine loads at C.R.=15. There is more reduction in CO emissions at higher compression ratio of 17.5 for both CNG substititION rates in dual fuel mode as compared to pure diesel mode as shown in figure 8.This reduction in CO emissions for duel fuel operation is due to the less injected diesel fuel and its relacement with a clean burning CNG fuel.

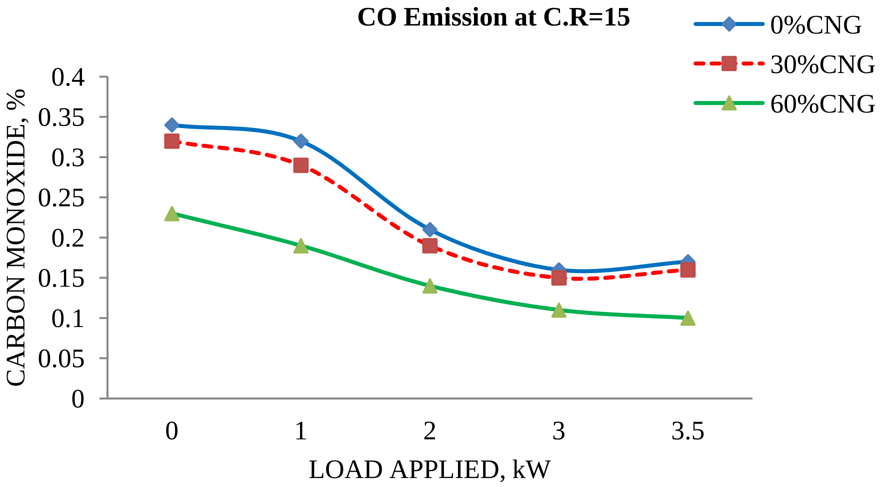


Fig.7: Carbon monoxide emission with load applied for diesel and dual fuel modes at C.R =15.

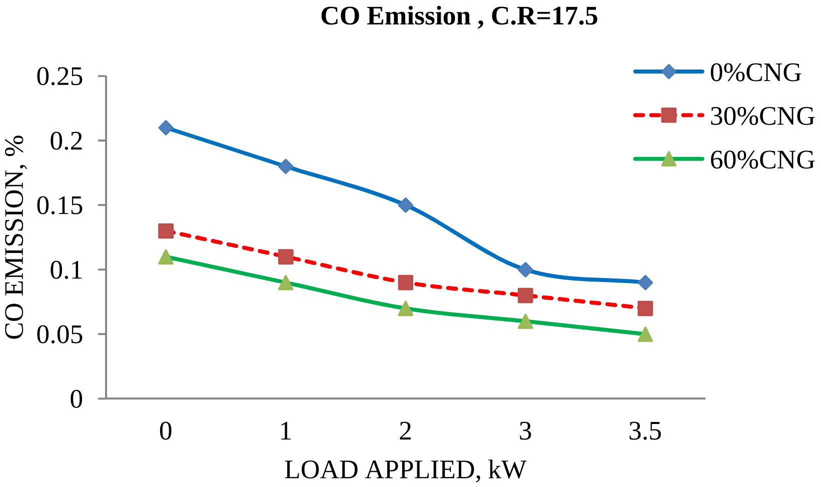


Fig.8: CO Emission with load applied for pure diesel and dual fuel modes at C.R=17.5.

4.2 HC Emissions analysis

Figures 9 and10 depict the variation of unburned HC emissions for diesel and dual fuel (with 30%CNG and 60%CNG) modes at the two compression ratios of 15 and 17.5 respectively. It is very clear from both the figures that HC concentration in the exhaust decrease with load applied for both diesel and dual fuel modes but

HC emissions are lesser for dual fuel mode than pure diesel mode by 14.55% at 1kW load and by 18.30% at 3.5kW load and by 28.16% at 1kW load and by 30.72% at 3.5kW load for 30% and 60% CNG substitution rates respectively at compression ratio of 15 as shown in figure 9. As the compression ratio is increased to 17.5, HC emissions are lower for dual fuel engine than diesel mode by 14.63% at 1kW load and 37.29% at 3.5kW load for 30% CNG substitution rate and by 17.07% at 1kW load and 44.92% at 3.5kW load for 60% CNG substitution rate as depicted in figure 10. With the use of small percentage of EGR in the engine cylinder in dual fuel mode, intake air temperature increased and as a result unburned hydrocarbon emissions in the engine exhaust decreased. At higher loads, higher compression ratio and higher CNG substitution rates, HC emissions further reduced for dual fuel operation because at these conditions the delay period decreases for pilot fuel and combustion of CNG-air mixture become fast and complete and very less amount of unburned fuel go into the engine exhaust.

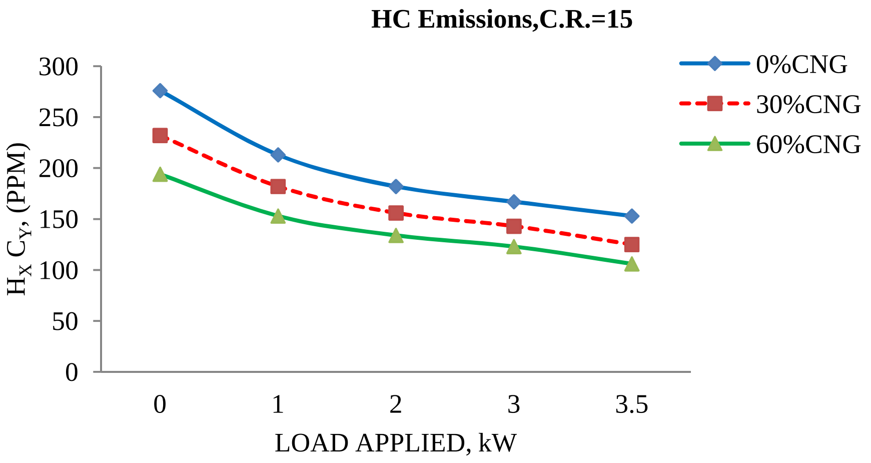


Fig.9: Hydrocarbon (ppm) emissions for diesel and dual fuel modes at C.R=15

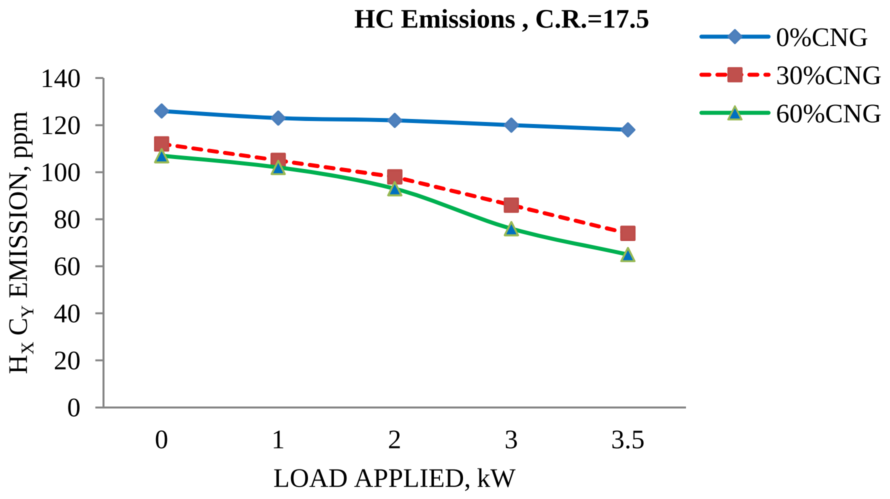


Fig.10: Hydrocarbon (ppm) emissions for diesel and dual fuel modes at C.R=17.5

4.3 CO_2 Emission analysis

Figures11 and 12 shows the effects of varying engine load on emission of carbon dioxide (CO_2). It can be observed that level of emission of CO_2 increase with increasing load both for diesel and dual fuel modes but its value for dual fuel mode is decreased by 17.24% at low load of 01kW and by 27.65% at rated load of 3.5 kW for C.R.=15 and 30% CNG substitution. At the same compression ratio, the concentration of these emissions further decrease by 44.83% for low load and by 59.57% at higher engine loads with 60% CNG substitution. CO_2 emissions further decrease with increase in compression ratio and CNG supply. This is evident from figure 12 that almost same trend is observed for 30% and 60% CNG dual fueling at higher compression ratio of 17.5. This reduction in levels of CO_2 emissions on dual fueling a converted diesel engine is beneficial in the sense that CO_2 is a greenhouse gas and its concentration in the atmosphere should be minimum. The main factors for reduction of CO_2 in exhaust of a dual fuel engine include improper conversion of CO to CO_2 due to decrease in peak temperature because of lower adiabatic flame temperature of CNG than diesel and decreased diesel fuel quantity.

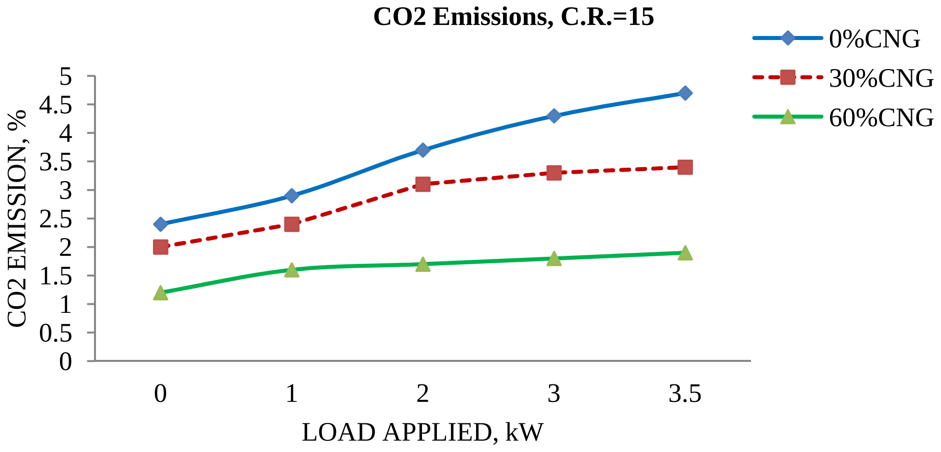


Fig. 11: CO_2 Emissions for diesel and dual fuel mode of operation at C.R. =15.

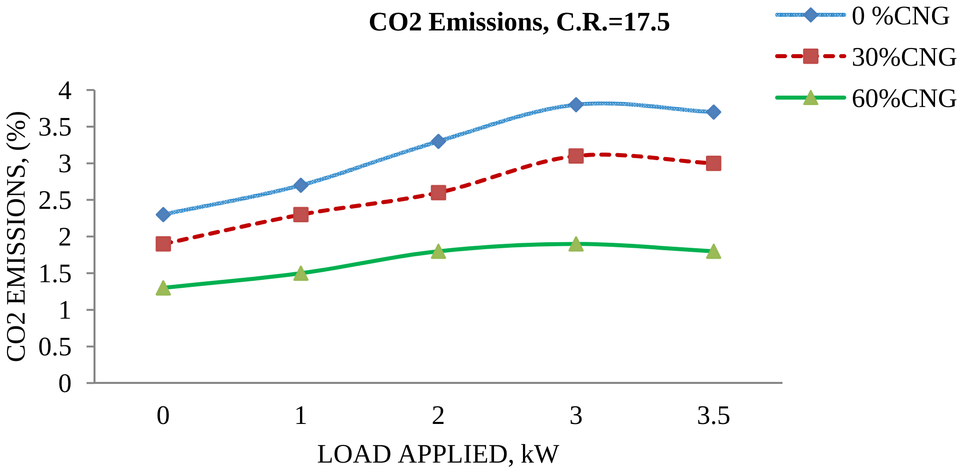


Fig.12: CO_2 Emissions for diesel and dual fuel mode of operation at C.R. =17.5.

4.4 NO_X Emissions analysis

With the introduction of gaseous fuels NO_X levels are found to be low. Figures 13 and 14 represent the effect of engine load, changing compression ratio and CNG substitution rates on NO_X emission formed inside engine cylinder for diesel and dual fuel modes. It is clear from figure 15 &16, that NO_X level increase with increase of engine load and compression ratio for both diesel and dual fuel modes but in dual fuel mode NO_X emissions are drastically reduced by 12.5% and 18.75% at low loads for 30% CNG and by 42.36% and 76.94% at high engine loads for 60%CNG at C.R. =15. NO_X concentration is further reduced by 43.78% and 77.83% at low loads for 30%CNG and 40.45% and 84.76% at rated load of 3.5kW for 60%CNG at C.R.=17.5 as depicted infigure14.

High peak temperatures and availability of oxygen are the two main factors for the formation of NO_X and it is directly related to adiabatic flame temperature. So as the CNG is introduced NO_X emissions decrease and as CNG supply is increased, NO_X further decrease. This decrease in NO_X in dual fuel engine is a positive merit in view of environmental concerns.

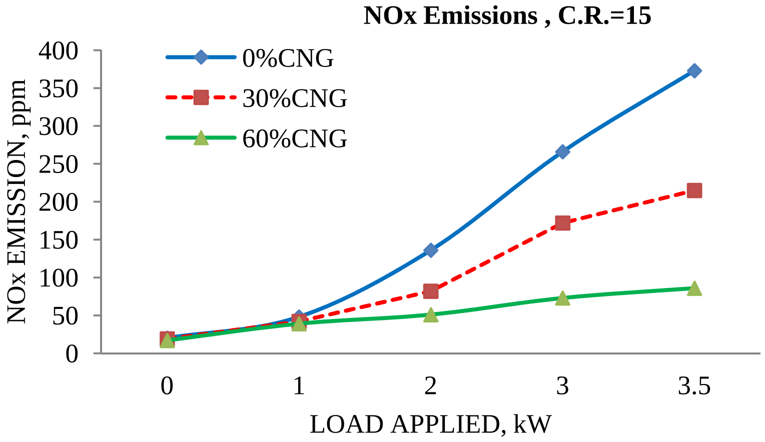


Fig.13: NO_X Emission as a function of engine load for diesel and dual fuel modes at C.R. =15.

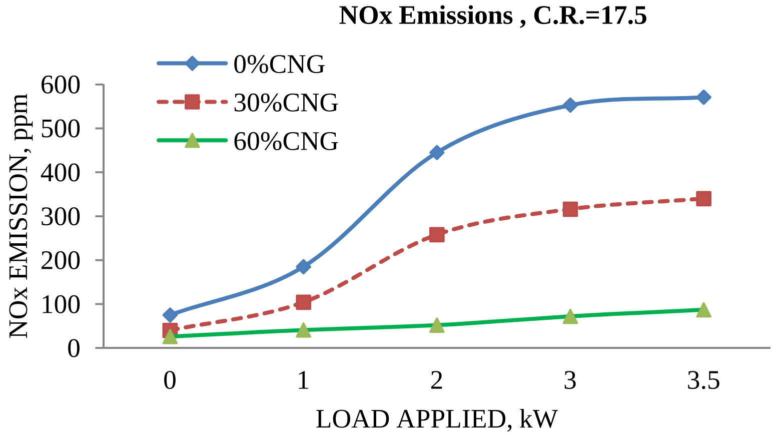


Fig.14: NO_X Emission as a function of engine load for diesel and dual fuel modes at C.R. =17.5.

4.5 Smoke opacity analysis

Smoke opacity means the degree to which the smoke reduces the passage of light. It means more smoke in the exhaust will have high smoke opacity and vice-versa in the context of diesel emissions. Figures 15 and 16, shows the effects of engine loads, variation of compression ratio and changes in CNG substitution rates on smoke opacity for diesel and dual fuel modes. Smoke opacity for pure diesel engine is inversely proportional to compression ratio i.e. if compression ratio increases smoke opacity decreases and if compression ratio decreases smoke opacity increases. Experimental results confirm this fact that its value for compression ratio of 15 is above 60% whereas for compression ratio of 17.5, its value decreases to 5.8% at the same engine load of 3.5kW. From figure 15, it can be seen that smoke opacity decreases by 25.61% at low loads and by 54.83% at rated load of 3.5kW for 30%CNG substitution and further decreases by 84.14% at low loads and 86.25% at the rated load of 3.5kW for 60% CNG substitution for compression ratio of 15. With increase in compression ratio from 15 to 17.5, results of smoke opacity are depicted in figure 16, and its value for 30%CNG reduces by 50% at low load and 84.48% at rated load of 3.5kW and it further decreases by 71.43% at low loads, 93.10% at 3.5kW load for 60%CNG supply in comparison to its values for pure diesel mode.

This decrease in smoke level with dual fueling the existing diesel engine is a positive merit in favor of dual fuel engines because diesel engines smoke reduction is the main cause of concern for researchers, manufacturers and users. The main factors of decrease in smoke emissions due to dual fueling of a diesel engine include, reduced injected diesel fuel, complete and smooth combustion of clean CNG fuel. In dual fuel engine, a flame front is formed by the ignition of small quantity of pilot fuel which sweeps the homogeneous mixture of CNG and air and exhaust contains less unburned fuel and hence less smoke. Moreover, soot particles form primarily from the carbon in the diesel fuel and in CNG, hydrogen/carbon ratio is high because of presence of smaller hydrocarbon as compared to diesel, soot formation is less and as a result Particulate Matter (PM) emission will also decrease with the use of CNG.

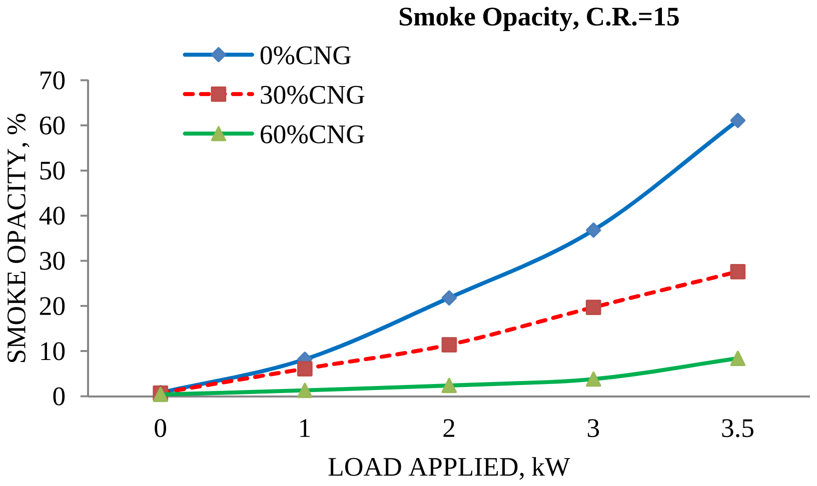


Fig. 15: Smoke opacity analysis for diesel and dual fuel modes at C.R. =15.

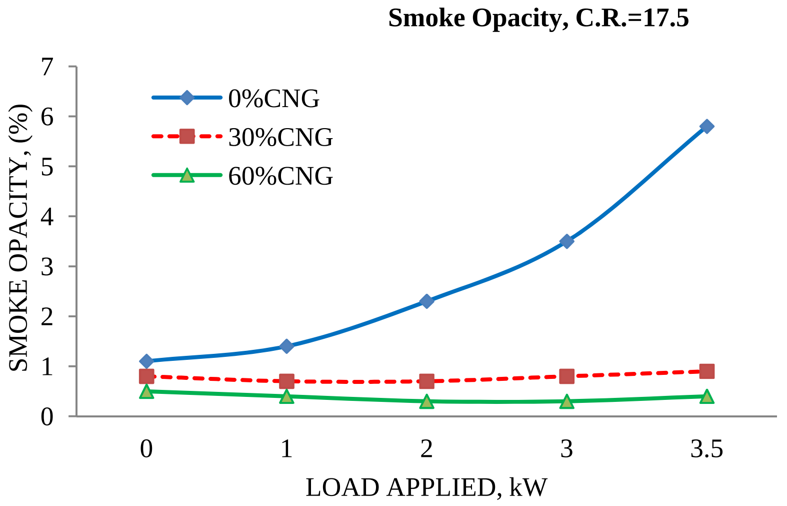


Fig. 16: Smoke opacity analysis for diesel and dual fuel modes at C.R. =17.5

5. CONCLUSION

In the present experimental research on performance and emissions on use of CNG in a converted direct injection conventional diesel engine in dual fuel mode, thorough investigations on two different compression ratios of 15 and 17.5 for 30% and 60%CNG substitution rates have been carried out at optimum operating conditions mentioned in table 3, the main conclusions are summarized below:

(1) Brake thermal efficiency of dual fuel mode with 30% CNG substitution rate, is more than diesel fuel mode by 5.11%, 5.58%, 9.77% and 10.74% at 1, 2, 3 and 3.5kW loads respectively. For 60%CNG substitution the value of brake thermal efficiency of dual fuel mode is more than neat diesel mode by 9.03%,11.17%,13.43% and 12.6% at 1, 2, 3 and 3.5kW engine loads respectively. For the two CNG substitution rates of 30% and 60%, brake thermal efficiency of dual fuel engine is more than that of pure diesel operation from no load to full load and this difference is maximum at full load and follow almost the same trend as at CR of 15 with slightly higher value at CR of 17.5.

(2) It is observed that BSFC with 30%CNG dual fuel mode is less than pure diesel mode by 59.49%, 69.04%, 57.14%, 61.54% at 1, 2, 3 and 3.5kW engine loads respectively at compression ratio of 15. The same trend is observed for 60%CNG substitution rate. As the compression ratio is further increased to 17.5, BSFC value for dual fuel mode is further decreased by 77.02%, 68.29%, 60.71% and 61.54% at 1, 2, 3, 3.5kW loads respectively at 30%CNG substitution rate in comparison to pure diesel mode. It further decreased as the substitution rate of CNG was increased to 60%.

(3) CO emissions for 30% and 60% CNG substitution rates are lower than pure diesel mode in the range of 33.3%-61.9% for different engine loads at C.R.=15. There is more reduction in CO emissions at higher compression ratio of 17.5 for both CNG substitition rates in dual fuel mode as compared to pure diesel mode.

(4) HC emissions are lesser for dual fuel mode than pure diesel mode by 14.55% at 1kW load and by 18.30% at 3.5kW load and by 28.16% at 1kW load and by 30.72% at 3.5kW load for 30% and 60% CNG substitution rates respectively at compression ratio of 15. As the compression ratio is increased to 17.5, HC emissions are lower for dual fuel engine than diesel mode by 14.63% at 1kW load and 37.29% at 3.5kW load for 30% CNG substitution rate and by 17.07% at 1kW load and 44.92% at 3.5kW load for 60% CNG substitution rate.

(5) CO_2 emissions increase with increasing load both for diesel and dual fuel modes but its value for dual fuel mode is decreased by 17.24% at low load of 01kW and by 27.65% at rated load of 3.5kW for C.R.=15 and 30% CNG substitution. At the same compression ratio, the concentration of these emissions further decrease by 44.83% for low load and by 59.57% at higher engine loads with 60% CNG substitution. CO_2 emissions further decrease with increase in compression ratio and CNG supply.

(6) In dual fuel mode NO_X emissions are drastically reduced by 12.5% and 18.75% at low loads for 30% CNG and by 42.36% and 76.94% at high engine loads for 60%CNG at C.R.=15. NO_X concentration is further reduced by 43.78% and 77.83% at low loads for 30%CNG and 40.45% and 84.76% at rated load of 3.5kW for 60%CNG at C.R.=17.5.

(7) Smoke opacity in dual fuel mode decreased by 25.61% at low loads and by 54.83% at rated load of 3.5kW for 30%CNG substitution and further decreases by 84.14% at low loads and 86.25% at the rated load of 3.5kW for 60% CNG substitution for compression ratio of 15. With increase in compression ratio from 15 to 17.5, its value for 30%CNG reduces by 50% at low load and 84.48% at rated load of 3.5kW and it further decreases by 71.43% at low loads, 93.10% at 3.5kW load for 60%CNG supply in comparison to its values for pure diesel mode.

In view of the above positive results obtained in this experimental research in favor of CNG-Diesel dual fuel engine on performance and emissions, it can be concluded that it is a promising technology for achieving better thermal efficiency and controlling both NO_X and smoke emissions in existing conventional compression ignition engines with minor engine hardware modifications, thus great saving of precious diesel fuel and saving the human and plant life from the hazardous effects of exhaust gas pollutants from the conventional diesel engines.

REFERNCES

[1] Bhandari, K et al. (2005): Performance and emissions of natural gas fueled internal combustion engine, A review, *JSIR Volume* **64** 333-338.

[2] Doughty, G E *et al.* (1992): Natural gas fuelling of a caterpillar 3406 diesel engine. *ASME* J *Engg gas turbine and power*.

[3] Guowei, L *et al*. (1999): Optimization study of pilot-ignited natural gas direct-injection in diesel engines, *SAE Paper No.* 1999-01-3556. 1739-1748

[4] Karim, G A (1983): The dual fuel engine of the compression ignition type –prospects, problems and solutions –a review. *SAE Paper No.* 831073.

[5] Karim, G A (1991): An examination of some measures for improving the performance of gas fuelled diesel engines at light load. *SAE Paper No.* 912366.

[6] Karim, G A and Burn, K S (1980): The combustion of gaseous fuels in a dual fuel engine of compression ignition type with particular reference to cold intake temperature conditions, *SAE Paper No.*800263.

[7] Liu, Z and Karim, G A (1997): Simulation of combustion process in gas-fuelled diesel engines. *Proc. Instn. Mech. Engrs. Volume,* **211** Part A, 159-169.

[8] Roydon, A. Fraser (1991): Auto ignition of methane and natural gas in a simulated diesel engine environment. *SAE Paper No*. 910227.

[9] Sera, M.A et al. (2003): CNG engine performance improvement strategy through advanced intake system, *SAE Paper No*. 2003-01-1937

[10] Singh, R and Sagar, M (2012), Dual fueling of a twin-cylinder compression ignition engine with diesel and CNG, *Journal of engineering and applied sciences*, **7,** 90-99

[11] Singh, S *et al.* (2004): Modeling and experiments of dual fuel engine combustion and emissions, *SAE Paper No*. 2004-01-0092

[12] Xianhua, D and Philip, H (1986): Emissions and fuel economy of a pre-chamber diesel engine with natural gas dual fueling, *SAE Paper No*. 860069.

[13] Youtong, Z *et al.* (2003): Modeling and simulation of a dual fuel (diesel/natural gas) engine with multidimensional CFD, *SAE Paper No*. 2003-01-0755, 336-346.

Performance Study of genus 3 Hyperelliptic Curve Cryptosystem

Daya Gupta*, Asok De** and Kakali Chatterjee*

Abstract—Hyperelliptic Curve Cryptosystem (HECC) is well suited for all kinds of embedded processor architectures, where resources such as storage, time, or power are constrained due to short operand sizes. We can construct genus 3 HECC on 54-bit finite fields in order to achieve the same security level as 160-bit ECC or 1024-bit RSA due to the algebraic structure of Hyperelliptic Curve. This paper explores various possible attacks to the discrete logarithm in the Jacobian of a Hyperelliptic Curve (HEC) and addition and doubling of the divisor using explicit formula to speed up the scalar multiplication. Our aim is to develop a cryptosystem that can sign and authenticate documents and encrypt / decrypt messages efficiently for constrained devices in wireless networks. The performance of our proposed cryptosystem is comparable with that of ECC and the security analysis shows that it can resist the major attacks in wireless networks.

Keywords— Hyperelliptic Curve Cryptosystem(HECC), Secure Hyperelliptic Curve, Hyperelliptic Curve Deffie-Hellman(HECDH), Hyperelliptic Curve Digital Signature Algorithm (HECDSA)

1. INTRODUCTION

Public Key Cryptography plays an essential role in wireless network as these networks are vulnerable to both active and passive attacks. Security mechanisms are essential to ensure the integrity, confidentiality, and authenticity of the data that are transmitted in such networks. A lot of information can be revealed online due to various attacks like forgery, impersonation attack, insertion attack, etc. Therefore, to protect the information between the user (client) and information provider (server), the message must be transmitted in an encrypted way. If an encrypted message is transmitted, the cryptographic framework has to ensure that no other party can obtain any information about the message or change it without being noticed. To fulfill this need, a digital signature is necessary which helps to guarantee the reliability, non-repudiation and unforgeability. Most of the systems use RSA based digital signatures [1, 2]. However RSA signature is not suitable for constrained devices as it requires long key length. This approach can lead to a number of problems such as increased processing time (decryption time increases about 8 times as key sizes double) and increased key storage requirement (for private and public key certificates). Cryptosystems based on Elliptic Curve and Hyperelliptic Curve are considered to

Manuscript received August 24, 2011; first revision November 30, 2011; accepted December 23, 2011.

Corresponding Author: Kakali Chatterjee

* Computer Engineering Department, Delhi Technological University, Bawana Road, Delhi 110042, India (daya_gupta2005 @yahoo.co.in, kakali2008@gmail.com)

** Ambedkar Institute of Advanced Communication Technologies & Research, Delhi 110031, India (asok.de@gmail.com)

be suitable for platforms with limited resources since they require smaller fields than RSA to attain the same security level. The purpose of this paper is the performance study of an alternative technology, namely HECC, which is based on curve arithmetic and offer significant benefits over RSA and ECC when used in constraint devices in wireless network.

To provide secure communication in a wireless network, authenticated key agreement protocol is an important primitive for establishing session key. Existing authentication protocol [3] certify users through a third party called Certification Authority who issues public keys for secure communication. But this increases the traffic a lot by introducing frequent certificates, which results in higher energy consumption and also key administration overhead. To overcome these problems, many key exchange protocols like PKE [4], EPA [5], and SKA [6] are introduced. Most of these protocols are password based where a directed dictionary attack can almost always succeed to break the password. To overcome this problem, ECC based authenticated key agreement protocol in wireless network are discussed in [7-10]. These protocols utilize ECDSA signature technique which enhances the security level of user authentication and key exchange. However, the security level can also be increased using hyperelliptic curve because it has some advantage over ECC. For instance, in ECC we have to work with operand lengths of approximately 160-bit whereas in the case of HECC, one needs 40-bit to 80-bit long operands to compute the group operations for these curves. Thus, HECC is more suitable for implementation in the constrained platforms like the PDA, smartcard, and handheld devices etc. in wireless network.

Current research on HECC emphasizes finding efficient methods to select secure hyperelliptic curves, fast operations on the Jacobians, and implementation of HECC for use in practical applications to enhance network security. In this direction, we have explored various possible attacks to the discrete logarithm in the Jacobian of a hyperelliptic curve that are to be considered to establish a secure HEC. Then we explore the group operations on a Jacobian in detail so as to obtain the explicit formula for performing addition and doubling in an efficient way to speed-up the arithmetic on genus-3 hyperelliptic curves.

Our contributions in this paper are as follows:

i) We have implemented Hyperelliptic Curve Diffie-Hellman (HECDH) key agreement protocol for secret key generation and Hyperelliptic Curve Digital Signature Algorithm (HECDSA) for signature generation /verification, considering the genus 3 Hyperelliptic Curve $C: v^2 + (u^2 + u)v = u^7 + u^5 + u^4 + u^3 + u^2 + u + 1$ over the finite field $\mathbb{F}_2{}^7$ using Netbeans IDE 6.8.
We have preferred Java over C++, as Java enables the development of robust applications on multiple platforms in heterogeneous distributed networks.

ii) The security analysis of the proposed protocol shows that it can resist the main attacks from both internal users and external hackers.

The rest of the paper is organized as follows:

In Section 2, Secure Hyperelliptic Curve is discussed; Section 3 provides Mathematical Background; Section 4 presents proposed Hyperelliptic Curve Cryptosystem; Section 5 presents Implementation Results; Section 6 presents Security Analysis of the proposed cryptosystem. Finally, we conclude the paper in Section 7.

2. Secure Hyperelliptic Curve

Hyperelliptic Curves present a rich source of abelian groups over which the discrete logarithm problem is believed to be difficult. Hence these groups can be used for the implementation of various public key primitives. If the curves are chosen carefully then the DLP in these groups is as hard as for general groups and one can use much smaller parameters and key sizes than when using the multiplicative group of finite fields and still obtain the same level of security. The selection of a secure hyperelliptic curve is explored in this section:

Let a hyperelliptic curve *C* of genus g is defined on a finite field $\mathbb{F}_q$ ($q = p^r$ and p is a prime), and given by the equation $C: y^2 + h(x)y = f(x)$. The order of the Jacobian $\mathbb{J}(\mathbb{F}_q)$ of C, denoted by # $\mathbb{J}(C; \mathbb{F}_q)$, should be divisible by a large prime number *l* of at least 40 decimal digits [11].

Let $\mathbb{F}_q{}^n$ denote the degree n extension of $\mathbb{F}_q$. Its Jacobian $\mathbb{J}(C; \mathbb{F}_q{}^n)$ over $\mathbb{F}_q{}^n$ is a finite abelian group and $(q^{n/2}-1)^{2g} \leq \#\mathbb{J}(C; \mathbb{F}_q{}^n) \leq (q^{n/2}+1)^{2g}$. The HCDLP in $\mathbb{J}(C; \mathbb{F}_q{}^n)$ is: given two divisors D_1, D_2 defined on $\mathbb{J}(C; \mathbb{F}_q{}^n)$ over $\mathbb{F}_q{}^n$, to determine the integer m such that $D_2=mD_1$, provided that such an integer *m* exists.

To establish a secure HEC, we should select the hyperelliptic curve so that its Jacobian satisfies the following conditions:

1) There is an index calculus attack on $\mathbb{J}(C; \mathbb{F}_q{}^n)$ that is more efficient than Pollard's rho method if the genus g of C is not small enough. Initially, this attack was developed for high genus curves by Adleman, Demarrais, and Huang [12]. They found a sub-exponent time algorithm to solve the DL in the Jacobian of hyperelliptic curves of a big genus over a finite field. Curves of higher genera (preferably $g \leq 4$) are, therefore, not suitable for cryptographic use ($2g+1 < \log q^n$).
2) Any naive implementation that neglects security aspects like selection of the curve can be broken by the generic cryptanalytic attacks. If the group order is large, but is divisible by only small primes, the DLP can be broken by Pohlig-Hellman attack. Therefore, $\#\mathbb{J}(C; \mathbb{F}_q{}^n)$ should have a large prime factor so as to prevent the attacks of Pohlig-Hellman's methods. Since the time complexity of Pohlig-Hellman's method is proportional to the square root of the largest prime factor of $\#\mathbb{J}(C; \mathbb{F}_q{}^n)$, so far it is demanded that this largest prime factor should be at least 160-bit in length.
3) In order to prevent the attack of Frey [13], which uses the Tate pairing generation of MOV attacks, the large prime factor of # $\mathbb{J}(C; \mathbb{F}_q{}^n)$ should not divide $(q^n)^k - 1$, here $k < (\log q^n)^2$.
4) In order to prevent the attack generated by Ruck [14], the Jacobian of a hyperelliptic curve over the large prime field GF(p) should not have p-order subgroup.
5) Simple side channel attacks obtain information from a single scalar multiplication by observing leaked information [15]. For restricted devices like smartcards it is possible for an attacker to derive side-channel information on the operations performed. To harden a cryptographic primitive against simple side-channel attacks, we make the observable information independent of the secret scalar. This is achieved by one of the following three approaches: inserting dummy arithmetic instructions, using indistinguishable / unified addition and doubling formulas, or applying Montgomery's ladder for scalar multiplication.

3. MATHEMATICAL BACKGROUND

Hyperelliptic Curve Cryptosystem (HECC) was proposed by Koblitz [11], based on the discrete logarithm problem on the Jacobian of hyperelliptic curves over finite fields.

3.1 Arithmetic of Hyperelliptic Curve

A hyperelliptic curve C of genus g over $\mathbb{F}_q$ is an absolutely irreducible non-singular curve defined by $C : y^2 + h(x)y = f(x)$, where $h, f \in \mathbb{F}_q$ [x], are such that $y^2 + h(x)y - f(x)$ is absolutely irreducible over $\mathbb{F}_q$, and if $b^2 + h(a)b = f(a)$, for (a,b) $\in \overline{\mathbb{F}} \times \overline{\mathbb{F}}$, then $2b + h(a) \neq 0$ or $h'(a)\, b - f'(a) \neq 0$ [16].

Definition 1- A hyperelliptic curve C is called an imaginary curve if q is odd, then f is monic, deg $(f) = 2g + 1$ and $h = 0$. If q is even, then h and f are monic, $deg(f) = 2g + 1$ and $deg\ (h) \leq g$.

Definition 2- A hyperelliptic curve C is called a real curve if q is odd, then f is monic, deg $(f) = 2g + 2$ and $h = 0$. If q is even, then h is monic , deg$(h) = g + 1$ and deg$(f) \leq 2g + 1$ or $deg(f) = 2g + 2$ and the leading coefficient of f is of the form $\beta^2 + \beta$ for some $\beta \in \mathbb{F}_q$.

Definition 3- A hyperelliptic curve C is called an unusual curve if $\mathbb{F}_q$ has odd characteristic, then $deg\ (f) = 2g + 2$ and if $\mathbb{F}_q$ has characteristic 2, then deg $(h) = g + 1$ and $deg\ (f) = 2g + 2$ and the leading coefficient of f is not of the form $\beta^2 + \beta$ for some $\beta \in \mathbb{F}_q$

Consider a Hyperelliptic curve C as defined by $C: y^2 + h(x)y = f(x)$. A divisor $D = \sum m_i P_i, m_i \in \mathbb{Z}$, is a finite formal sum of $\overline{\mathbb{F}}$ points. Its degree is the sum of the coefficients $\sum m_i$. The set of all divisors form an Abelian group denoted by $\mathbb{D}$(C). The set of degree zero divisors $\mathbb{D}^0$ forms a subgroup of $\mathbb{D}$(C).

Every rational function consisting of the formal sum of the poles and zeros of the function on the curve C gives rise to a divisor of degree zero. Such divisors are called principal and the set of all principal divisors is denoted by $\mathbb{P}$. If $D_1, D_2 \in \mathbb{D}^0$ then we write $D_1 \sim D_2$ if $D_1 - D_2 \in \mathbb{P}$; D_1 and D_2 are said to be equivalent divisors. Now, we can define the Jacobian of C as the quotient group $\mathbb{D}^0/\mathbb{P}$ [17].

If we want to define the Jacobian over $\mathbb{F}$, denoted by $\mathbb{J}_C(\mathbb{F})$, we say that a divisor $D = \sum m_i P_i$ is defined over $\mathbb{F}$ if $D^\sigma = \sum m_i P_i^\sigma$ is equal to D for all automorphisms σ of $\overline{\mathbb{F}}$ over $\mathbb{F}$ [16].

Cantor shows that each element of the Jacobian can be represented in the form $D = \sum^r_{i=1} P_i - r.\propto$ such that for all $i \neq j$, P_i and P_j are not symmetric points [18]. Such a divisor is called a semi-reduced divisor. Each element of the Jacobian can be represented uniquely by such a divisor, subject to $r \leq g$. Such divisors are referred to as reduced divisors. We use the reduced divisor in addition to $\mathbb{J}_C$.

For a genus 3 hyperelliptic curve C over $\mathbb{F}_q$ defined as C: $Y^2 = \mathbb{F}(X)$, a semi-reduced divisor can be represented by the following pair of polynomials :

(**Mumford's representation**)
$D = (U, V), U,V \in \mathbb{F}_q[X]$; where
$U = \prod(X - x_i)^{ordpi\ (D)}$,
$y_i = V(x_i)$

for $P_i = (x_i,y_i) \in C$ with $\mathrm{ord}_{pi}(D) > 0$, and
deg V < deg U, $F - V^2 \equiv 0 \bmod U$.

The degree of U is called the weight of D, and D is a reduced divisor, if its weight equals 3. Any class in $J_C(\mathbb{F}_q)$ is uniquely represented by a reduced divisor (i.e., each class includes a unique reduced divisor).

Unlike elliptic curve, the points on the hyperelliptic curve do not form a group. The additive group on which the cryptographic primitives are implemented is the divisor class group. Each element of this group is a reduced divisor. The group elements have a nice cannonical representation by means of two polynomials of small degree. The basic algorithm of performing arithmetic for divisor addition and doubling in the Jacobian of hyperelliptic curve is Cantor's algorithms.

3.2 Group Operations on a Jacobian

Group operations of divisor on $\mathbb{J}_c(\mathbb{F})$ are performed in the following two steps: addition of generic divisors and doubling of generic divisors. Addition of divisor classes means multiplication of ideal classes, which consists of a composition of the ideals and a first reduction to a basis of two polynomials. The output of this algorithm is called semi-reduced divisor. Then the second algorithm (reduction) is used to find the unique representative in the class. Cantor’s algorithm [18] is used for transferring the group laws in a sequence of composition and reduction using only polynomial arithmetic.

```
Cantor’s Algorithm for Group Addition
Input: $D_1 = [U_1, V_1]$ and $D_2 = [U_2, V_2]$ , $C:Y^2 = F(X)$
Output: $D_3=(U_3,V_3)$ reduced with $D_3 = D_1+ D_2$
1. Compute $d_1 = \gcd(u_1, u_2) = e_1u_1 + e_2u_2$;
2. Compute $d = \gcd(d_1,v_1+v_2+h) = c_1d_1 + c_2(v_1+v_2+h)$;
3. Let $s_1 = c_1e_1$, $s_2= c_1e_2$, $s_3 = c_2$;
4. $U' = u_1u_2/d^2$;
5. $V' = \{s_1u_1v_2+s_2u_2v_1+s_3(v_1v_2+f)\}/d \bmod u$
6. $U_3= (F-v^2) / U'$, $V_3= - V' \bmod U_3$
7. Make $U_3$ monic.
```

Harley’s algorithm for genus 3 Curve: Cantor’s algorithm is slow due to Polynomial arithmetic. The solution is to transform polynomial operations into field operations (explicit formula) by considering most frequent cases (occur with a probability ~1- O (1/q)). It was done by Harley in 2000 [19] by using reduced divisors represented by Mumford’s representation for input and output divisor classes on genus 2 curves. Subsequently, Harley’s algorithm was used for fast arithmetic on genus 3 HEC.

First classification of the input divisor classes is done by using the weights of them and then another classification of the divisor classes is done by testing $\gcd(U_1,U_2) = 1$ for addition $D_3 = D_1 + D_2$, $D_1 = (U_1,V_1)$, $D_2 = (U_2, V_2)$ or by testing $\gcd(U_1,2V_1) = 1$ for doubling $D_2 = 2D_1$, $D_1 = (U_1, V_1)$. These gcd computations are carried out by a resultant computation. The case satisfied deg U_1 = deg U_2 = 3 and $\gcd(U_1,U_2) = 1$ for addition and the case satisfied deg $U_1 = 3$ and $\gcd(U_1,2V_1) = 1$ for doubling, are called the most frequent cases.

3.3 Explicit Formula for genus 3 Curves

The first explicit formula for genus 2 proposed by Harley [19] has been followed by the work of Lange [20, 21]. After extensive research on explicit formula for performing addition and doubling, Avanzi [22] proposes a software implementation of genus 2 and genus 3 hyperelliptic curves over large prime fields. Pelzl and Wollinger [17, 23] propose a cost effective explicit formula for genus 2 and genus 3 curves and give the first implementation of a HEC cryptosystem on an embedded processor. Gonda et al. [24] propose improvements of addition algorithm on genus 3 HEC and implemented it on a 64-bit CPU. Fan et al. [25] discussed the performance of genus 3 HECC over three different binary fields. The idea to use HEC for cryptographic applications has been further analyzed and implemented in software and hardware oriented platforms by Kuroki et al. [26], Sakai and Sakurai [27], Nagao [28], and Smith [29]. We discussed the evolution of Hyperelliptic Curve Cryptosystems in [30].

Explicit Formula for addition and doubling of divisors based on [17, 23, and 25] is discussed below. The most frequent input for addition consists of two divisor classes represented by $[U_1,V_1]$, $[U_2,V_2]$, where $\deg(U_1) = \deg(U_2) = 3$ and $\gcd(U_1,U_2) = 1$. This guarantees that the associated reduced divisors D_1, D_2 do not have any point or its opposite in common and both divisors have 3 affine points in the support. For the doubling it is assumed that the class is represented by $[U_1, V_1]$ with $\deg(U_1) = 3$ and that $\gcd(U_1, 2V_1) = 1$.

Explicit Formula for addition of divisors of genus 3 HEC (most frequent cases)

Input:- C: $Y^2 = F(x)$, $F=x^7+f_5x^5+f_4x^4+f_3x^3+f_2x^2+f_1x+f_0$;

Reduced divisors $D_1= (U_1,V_1)$ and $D_2= (U_2,V_2)$

$U_1=x^3+u_{12}x^2+u_{11}x+u_{10}$, $V_1=v_{12}x^2+v_{11}x+v_{10}$, $U_2=x^3+u_{22}x^2+u_{21}x+u_{20}$, $V_2= v_{22}x^2+v_{21}x+v_{20}$

Output:- Reduced divisor $D_3=(U_3,V_3)=D_1+D_2$

$U_3 = x^3+u_{32}x^2+u_{31}x+u_{30}$, $V_3= v_{32}x^2+ v_{31}x+v_{30}$

```
Steps:- 1. Compute Resultant r of $U_1$ and $U_2$. [using Bezout's theorem]
  If r=0 then call the Cantor Algorithm.
2. Compute pseudo inverse $I=i_2x^2+i_1x+i_0 \equiv r/U_1 \bmod U_2$
3. Compute $S'=s'_2x^2+s'_1x+s_0' = rS \equiv (V_2-V_1)I \bmod U_2$ [using Karatsuba multiplication]
4. Compute $S= (S'/r)$ and make S monic. [using Montgomery trick]
5. Compute $Z=x^5+z_4x^4+z_3x^3+z_2x^2+z_1x+z_0=SU_1$[using Karatsuba multiplication]
6. Computing $U_t = x^4+u_{t3}x^3+u_{t2}x^2+u_{t1}x+u_{t0} = (S(Z+2w_iV_1) - w_i^2((F-V_1^2)/U_1))/U_2$
   [using karatsuba multiplication, Efficient Division]
7. Compute $V_t = v_{t3}x^3 + v_{t2}x^2+v_{t1}x+v_{t0} \equiv -(wZ+ V_1) \bmod U_t$
8. Compute $U_3=x^3+u_{32}x^2+u_{31}x+u_{30} = (F-V_t^2)/U_t$ [using Efficient Division]
9. Compute $V_3=v_{32}x^2+v_{31}x+v_{30} \equiv - V_t \bmod U_3$
```

Explicit Formula for doubling of divisor of genus 3 HEC (most frequent cases)

Input:- C: $Y^2 =F(x)$, $F= x^7+f_5x^5+f_4x^4+f_3x^3+f_2x^2+f_1x+f_0$;

Reduced Divisor $D_1= (U_1, V_1)$, $U_1 = x^3+u_{12}x^2+u_{11}x+u_{10}$, $V_1= v_{12}x^2+v_{11}x+v_{10}$

Output:- Reduced divisor $D_2=(U_2,V_2)=2D_1$, $U_2= x^3+u_{22}x^2+u_{21}x+u_{20}$, $V_2= v_{22}x^2+v_{21}x+v_{20}$

```
Steps:- 1. Compute Resultant r of $U_1$ and $2V_1$. [using Bezout's theorem]
  If r=0 then call the Cantor Algorithm.
2. Compute pseudo inverse $I = i_2x^2+i_1x+i_0 \equiv r/(2V_1) \bmod U_1$
```

3. Compute $Z=z_2x^2+z_1x+z_0=((F-V_1^2)/U_1)\mathrm{mod}U_1$ [using Efficient Division]
4. Compute $S' = s_2'x^2+s_1'x+s_0' = rS \equiv Z\ I\ \mathrm{mod}U_1$ [using Karatsuba multiplication]
5. Compute S=(S'/r) and make S monic [using Montgomery trick]
6. Compute $G=x^5+g_4x^4+g_3x^3+g_2x^2+g_1x+g_0=SU_1$ [using Karatsuba multiplication]
7. Compute $U_t = x^4+u_{t3}\,x^3+u_{t2}x^2+u_{t1}x +u_{t0} = ((G+w_iV_1)^2-w_i^2F)/U_1^2$
8. Compute $V_t= v_{t3}x^3+v_{t2}x^2+v_{t1}x+v_{t0}\equiv - (wG+V_1) \bmod U_t$
9. Compute $U_2= x^3+u_{22}x^2+u_{21}x+u_{20} = (F-V_t^2)/U_t$ [using Efficient Division]
10. Compute $V_2=v_{22}x^2+v_{21}x+v_{20} \equiv - V_t\ \mathrm{mod}U_2$

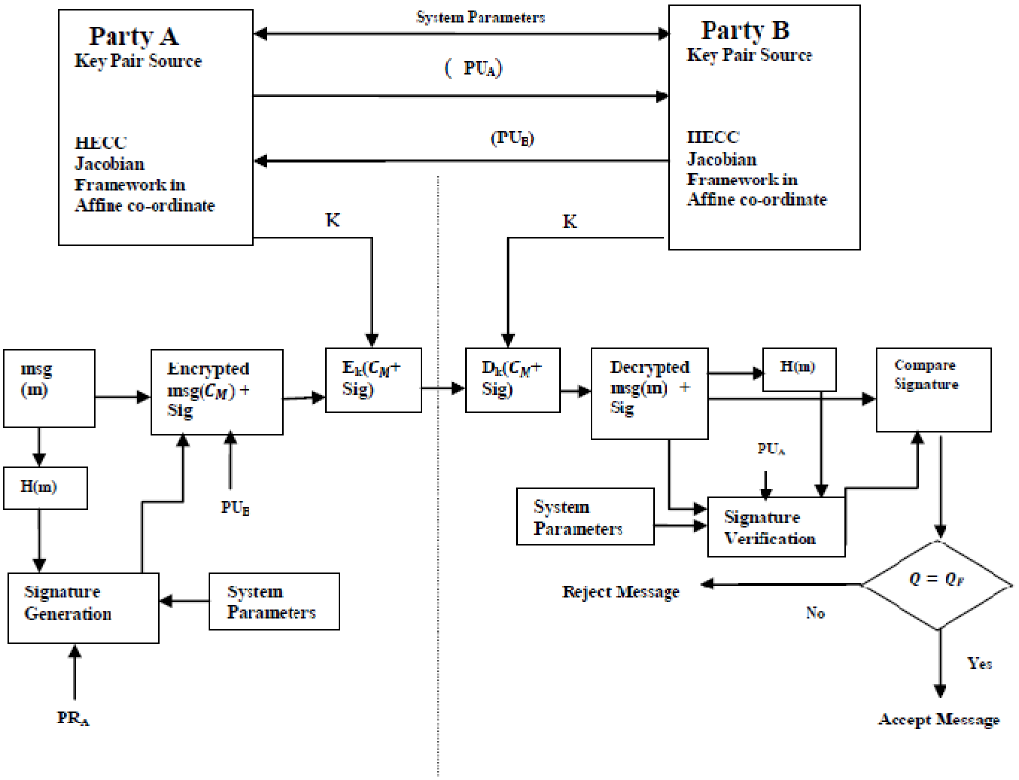
In the above case the cost of addition and doubling of divisor is I+68M+12S and I+60M+16S respectively where I= Inversion, M = Multiplication, S= Squaring.

4. Proposed Hyperelliptic Curve Cryptosystem

The proposed cryptosystem is shown in Fig.1.

The system works as follows:

Party A (Client) and Party B (Server) will generate their private keys (PR_A , PR_B) and public keys (PU_A, PU_B) using HEC. Before generating key pairs, the system parameters hyperelliptic curve C, prime p, divisor D, and elements of the group G of order N are exchanged in the initialization phase. For authentication, Party A first takes message *m*. The message is input to a hash function that produces a secure hash code of a fixed length. The hash code is provided as



input to a signature function along with a random number k generated by PRNG (Pseudo Random Number Generator) for this particular signature. The function also depends on PR_A and system parameters ($\mathbb{F}_2^n$, C , D, N, c, p). The result is a signature consisting of two components, Q and S. After that the message is first encrypted and binded with the signature. Then the block containing the encrypted message (C_M) and signature (Sig) is again encrypted with the common secret key (K) and send it to the Party B.

At the receiving end, Party B first decrypts the block containing the message and the signature with the common secret key K. Now Party B gets the encrypted message and signature separately. Again Party B will decrypt the encrypted message (C_M) and a hash code of the incoming message is generated. The incoming message and the signature is the input to a verification function. The verification function also depends on the PU_A, system parameters ($\mathbb{F}_2^n$,C,D,N,c, p). The output of the verification function is a value that is compared with the signature component. If the condition $(Q = Q_F)$ is satisfied then the signature is considered to be a valid signature and the message is accepted. We divide the proposed protocol into four different phases. Our proposed protocol uses modified HECDH and HECDSA to enhance security.

The phases of our proposed protocol are described below:

Initialization Phase

During this phase, Party A first calculates X= h(CIN) where CIN is the Client Identification No. and h( ) is a one way hash function. Now the client (Party A) generates a pre-knowledge message containing the client network identity (ID), X, and system parameters (hyperelliptic curve C, prime p and divisor D, representation of field elements of order N, and P a point on the curve) and transmit it to the server (Party B). Once the pre-knowledge message is received, server (Party B) will match the received value of X with the stored value in the verification table (this table is generated after the deployment of the network). If the values are same, then the server (Party B) agrees for further communication and sends an acknowledgement of the pre-knowledge message to Party A. After that both parties will generate a common secret key for secure communication.

Secret Key Generation Phase

In this phase we use Hyperelliptic Deffie-Hellman key exchange protocol for establishing a common secret key by applying the following steps:

- Party A will generate Private key $PR_A \in_R N$ [choose a prime (PR_A) at random in N].
- Party A will generate Public key $PU_A = [PR_A]\ D$ and send PU_A to Party B. [PU_A is represented using Mumford representation which is of the form $(u(x);\ v(x))$];
- Similarly Party B will generate the private key PR_B, public key $PU_B = [PR_B]D$. Party B will send the public key PU_B Party A.
- Once Party A and Party B exchange their public keys, Party A computes $M_1 = [PR_A]$ P and common secret key $K = (\ PR_A + M_1\ )(PU_B)$
  Here P is a point on the curve whose x co-ordinate is considered for calculation of $M_1,\ M_2$.
- On the other side Party B will computes $M_2 = [PR_B]$ P and generates the common secret key $K = (\ PR_B + M_2\ )(PU_A)$. If the protocol works correctly, both the Party A and Party B generates the same value of K. This can be proved by the simple mathematical calculation shown below:

$$K = (\,PR_A + M_1\,)(PU_B) = (\,PR_A + M_1\,)\,[PR_B]D = [PR_A][PR_B]D + [PR_A]\ \mathrm{P}\ [PR_B]D =$$
$$[PR_A]D\ [PR_B] + [PR_A]D\ [PR_B]\ \mathrm{P} = PU_A(\,PR_B + M_2\,)$$

Signature Generation and Verification Phase

HECDSA signature scheme is used here. For generation of signature we apply hash functions to message m. Here, we assume that E∈G is represented using Mumford representation which is of the form E = $[u_E, v_E]$ with u_E= x^e + $\sum_{i=0}^{e-1} u_i$ where $e \le$ g

► A generates a random nonce k $\in_R$ N and produce $E \leftarrow [k]D$ where D is each non-trival group element represented using Mumford representation.

► Party A calculates Q = $\sum_{i=0}^{e-1} \mathrm{L}(u_i)\mathrm{q}^{\mathrm{i}} \mathrm{mod\ p}$ where e is an integer with $e \le$ g and assuming the finite field elements are ordered such that 0 ≤ L(u$_i$) < q

► $S \leftarrow (k^{-1}(h(m) - [PR_A]Q))$ mod l where l =N/c (c is co-factor)

► The signature (Q, S) then binds with the message to provide authentication and send it to Party B.

For verification Party B follow the steps as described below:

► If Q or $S \notin [1, l-1]$ then reject the signature

► Else Party B generate $w \leftarrow S^{-1} mod\ l\ and$ R_1 and R_2 such that

► $R_1 \leftarrow [h(m)w]\ mod\ l$

► $R_2 \leftarrow [Qw]\ mod\ l$

► $F \leftarrow [\,R_1]\ D \oplus [\,R_2]\ PU_A$ where F is also in the form $F\ =\ [u_F, v_F]$

► If $F\ =\ 0$ then reject else calculate $Q_F\ =\ \sum_{i=0}^{e-1} \mathrm{L}(u_{F,i})\mathrm{q}^{\mathrm{i}} \mathrm{mod\ p}$

► If $Q\ =\ Q_F$ then accept the signature else reject.

Encryption and Decryption Phase

For message encryption and decryption, we follow the methods of HECC encryption and decryption [15].

In the encryption process first message m will be encoded as a series of points which is represented as $(u(x), v(x))$ noted as E$_M$. To encrypt this message, Party A will perform the following steps:

► Party A generates $W \leftarrow [k]D$ [W is in the form $(u(x);\ v(x))$] where k is previously generated random prime no. and D is the divisor.

► Produce the cipher text $C_M \leftarrow \{W, E_M + [k]PU_B\}$ which is sent to Party B.

To decrypt the cipher text C_M, Party B multiplies the first part in the pair by Party B's private key PR_B and subtracts the result from the second part of the pair. The original message can be retrieved from the cipher text as shown below:

$$E_M + [k]PU_B - PR_B\,(W)\ =\ E_M + [k]PU_B\ -\ PR_B\,[k]D\ =\ E_M + [k]PU_B - [k](PR_B\,D)$$
$$=\ E_M + [k]PU_\mathrm{B} - [k]PU_B\ =\ E_M$$

5. Implementation Results

We have implemented genus 3 HECC on different binary fields using Netbeans IDE 6.8. Netbeans refers to both a platform framework for Java desktop application and an Integrated

Table 1. Experimental Results of genus 3 HECC (Binary Field)

G-3 Binary field	**Curve (C) Equation**	**Divisor Generation**	**Public Key** PU_A	**Public Key** PU_B	**Party-A Secret key K**	**Party-B Secret key K**	**Signature Generation (Party-A)**	**File encryption (Party-A)**	**File decryption (Party-A)**	**Signature Verification (Party-A)**
Field Order-F(2^56) Group Order-(2^168)	C : v^2 + (u^2 + u)v = u^7 + u^5 + u^4 + u^3 + u^2 + u + 1 (2 ms)	D_1: div (u^3 + 12u^2 + 12u + 1, u^2 + 2u + 8) D_2: div (u^3 + u^2 + 13u +12, 12u^2+fu+ e) $D=D_1+D_2$: div (24u^3 + 1fu^2 + 1fu + f, 11u^2 + 1au + 13) Dinv: div (u^3 + 12u^2 + 12u + 1, 3u + 8) (6 ms)	div (u^3 + 14u^2 + u + 1f, fu^2 + u + 1) (14 ms)	div (u^3 + 16u^2 + 11, 7u^2 + 15u + 1d) (15 ms)	div (24u^3 + u^2 + 18u + 6, 6u^2 + 12u + 12) (14 ms)	div (24u^3 + u^2 + 18u + 6, 6u^2 + 12u + 12) (15 ms)	Signature = [AC0CFQCSOUYUNUJHldm0utAF6aCB5z,rawIUcQ6WYQJukWovro9uwY59gHGK5] Time in ms: 60	Encryption done : plain.txt.enc (580ms)	Decryption done : plain.txt.enc.txt (150 ms)	Signature verification successful! Time in ms: 30
Field Order-F(2^54) Group Order-(2^162)	C: v^2 + (u^2 + u)v = u^7 + u^5 + u^4 + u^3 + u^2 + u + 1 (2 ms)	D_1: div (u^3 + 11u^2 + 12u + 1, u^2 + 2u + 6) D_2: div (u^3 + u^2 + 13u + 12, 12u^2 + fu + e) $D=D_1+D_2$: div (u^3 + 1du^2 + 8u + 1, 10u^2 + 11u + 5) Dinv: div (u^3 + 11u^2 + 12u + 1, 3u + 6) (7 ms)	div (u^3 + 18u^2 + 1au + 12, au^2 + 4u + 1a) (16 ms)	div (24u^3 + 18u^2 + 19u + 15, 3u^2 + 17u + 8) (17 ms)	div (u^3 + 17u^2 + 12u + 4, 2u^2 + 7) (16 ms)	div (u^3 + 17u^2 + 12u + 4, 2u^2 + 7) (17 ms)	Signature = [FBACFEw4DEOhMZY5PBa8ertdfgEtDO,ym3JAhRbIF8OQIq6bfbYDszvZ4i4yCsg3Fw] Time in ms: 53	Encryption done : plain.txt.enc (682 ms)	Decryption done : plain.txt.enc.txt (180 ms)	Signature verification successful! Time in ms: 31

Development Environment (IDE) for developing mobile and web applications with Java. In the Java architecture, the security API (java.security package) for the Java Development Kit (JDK) introduced the Java Cryptography Architecture (JCA), which allows the generation of digital signature and message digests and more specifically Java Cryptography Extension (JCE), which provides implementation for key generation and agreement, encryption, and decryption algorithm.

We have considered the hyperelliptic curve $C: v^2 + (u^2 + u)v = u^7 + u^5 + u^4 + u^3 + u^2 + u$ +1 of genus 3 over the finite field $\mathbb{F}_{2^7}$. P=(α^{30} , 0) is an ordinary point in C($\mathbb{F}_{2^7}$). A divisor D can be represented as $D = div(a, b)$ with $a, b \in \mathbb{F}$, such that $D \sim D_1 + D_2$ where $D_1 = div(a_1, b_1)$, $D_2 = div(a_2, b_2)$. After computing the semi-reduced divisor $D = D_1 + D_2$ and $D' \sim D$, we have implemented Hyperelliptic Curve Digital Signature Algorithm in Netbeans IDE 6.8. The timings of basic operations using Netbeans IDE 6.8 are shown in Table 1 (after the first round). The timings have been measured on a PC with an Intel Core 2DUO CPU T6400@ 2.00GHz and windows vista operating system having jdk1.

5.1 Performance Analysis

i) We have compared our results with the timings of the basic operation found in [31]. These are 80 ms for secret key generation, 150 ms for signature generation, and 230 ms for signature verification with 163-bit key size based on ECC as against our result of 16 ms for secret key generation, 53 ms for signature generation, and 31 ms for signature verification

Table 2. Comparison of HECC with ECC for equivalent key sizes

Reference	Curve (key size)	Point Scalar Multiplication	Secret Key Encryption	Hash Function
Our result	HEC (54-bit key)	2	1	1
Lim et al. [10]	EC (160-bit key)	3	1	2
Aydos et al. [3]	EC (160-bit key)	3	2	1

with 54-bit key size based on HECC. Hence, our experimental result shows that the proposed cryptosystem is efficient as it takes less time for basic authentication operations.

ii) An efficient authentication protocol takes into consideration the communication and computation load during the user (client) authentication phase. The total number of bits exchanged in this protocol is 780 bits. The proposed protocol also has a low computation load on the client side compared to the existing protocols as shown in Table 2.

6. Security Analysis

In this section, we discuss the security of HECC. The proposed key agreement protocol will be considered to be a secure authenticated protocol if it satisfies the following properties:

Man-in-the-middle attack: This can be considered as an active attack. In this protocol, no useful information about the secret key K is revealed during a successful run. Consider A and B to be two communicating parties. Attacker I intercepts the public key PU_A and replaces it by PU_I. Then it sends PU_I to B and when B sends its public key PU_B, it again captures PU_B. But attacker I cannot compute the value of K and D because the security of an HEC is based on the difficulty of solving the discrete logarithm problem in the Jacobian of the curve. Thus attacker I cannot decrypt the useful messages or generate a valid signature during a successful run. Thus this protocol resists the man-in-the-middle attack.

Small subgroup attack: In a hyperelliptic cryptosystem, the system parameters ($\mathbb{F}_2{}^n$,C,D,N,c, p) are chosen in a manner such that the DL problem is too hard to compute. If hyperelliptic curve C has enough prime factors, the attacker could determine the secret scalar modulo of all these primes and recover a large part of the secret by using Chinese remaindering. This type of attack is called small subgroup attack. To avoid this attack, we check that D has order l where l is prime. For checking this, we first check that [l]D=0 and computing [h]D for $h=c/p_i$, for all prime divisors p_i of c and check that the result is not zero [15].

Known-key attack: In our proposed protocol, the client and the server both generate new PU_A and PU_B in every new session, and in addition the secret random no. k is changed with every new session also. Thus our proposed protocol is secure against known key attacks assuming that the hyperelliptic curve discrete logarithm problem is intractable.

Cipher text only attack: In a chosen ciphertext attack, the adversary knows the encryption algorithm and has access to many ciphertexts to be decrypted with an unknown key. In a chosen ciphertext attack the adversary uses the previous results to select subsequent ciphertexts with the secret key. In our proposed protocol, we never transmit the secret key K, which is used to encrypt the original encrypted message along with the signature. Even if an adversary knows the secret key, he cannot produce the original message as the cipher text consists of two parts, which is very difficult to decrypt as it lies on DLP in Jacobian of the curve.

Dictionary attack: In dictionary attack, the attacker pretends to be a legitimate client and attempts to login by guessing different passwords from a dictionary. As our protocol is not based on passwords, this type of attack is not applicable.

Perfect forward secrecy: In our protocol, perfect forward secrecy is maintained even if the user's public key is compromised since private key cannot be compromised as it is a random value that is changed from time to time. The adversary cannot decrypt the ciphertext as encryption is done using private key and public key, which depends on the difficulty of solving the discrete logarithm problem in the Jacobian of the curve. If we change the system parameters after each session then this problem becomes more hard. Thus, the property of perfect forward secrecy is satisfied.

7. Conclusion

HECC is a cryptosystem of choice when targeting embedded environments as the HEC operand size is only a fractional amount of the EC operand size and almost all the standard discrete logarithm based protocols such as the Digital Signature Algorithm (DSA) and EIGamal can be planted to HEC. We have explored in this paper various possible attacks that are to be considered to establish a secure HEC and efficient scalar multiplication. Our experimental results show that a public-key cryptosystem for constrained devices based on hyperelliptic curves can be designed to exchange keys, sign and authenticate documents, and encrypt and decrypt messages efficiently. In our view, HECC of genus 3 has the merit to be the preferred cryptosystem in constrained environment, as the performance of HECC with the operand size of 54-bit is comparable with the performance of ECC with an operand size of 160-bit.

References

[1] O.Goldreich, Y.Lindell, "*Session-Key Generation Using Human Passwords only," Crypto 2001, LNCS 2139*, pp.408-432.

[2] R.Katz, Q. Trovsky, M.Yang, *"Efficient Password Authenticated Key Exchange Using Human Memorable Passwords," Eurocrypt 2001, LNCS 2045*, pp.475-494.

[3] M.Aydos, T.Yanık, C.K.Koc, "*High-Speed Implementation of an ECC-based Wireless Authentication Protocol on an ARM Microprocessor", IEE Proceedings: Communications,* 2001, 148(5): pp.273-279.

[4] V. Boyko, P. Mackenzie, S. Patel. "*Provably secure password authenticated Key Exchange using Diffie-Hellman". EuroCrypt 2000, LNCS* pp.156-171.

[5] Y H Hwang, D H Yum, P J Lee, "*EPA: An Efficient Password-Based Protocol for Authenticated Key Exchange", ACISP 2003, LNCS 2727*, pp.452-463.

[6] E. Ryu, K. Kim, K. Yoo. "*A Simple Key Agreement Protocol", In Proc. of IEEE 37th Annual International Carnahan Conference 2003*, pp 128-131.

[7] K.Jung, J.Kim, T.Chung, "*Password-Based Independent Authentication and Key Exchange Protocol", ICICS-PCM 2003, IEEE*, pp.1908-1912.

[8] Julien Bringer, Hervé Chabanne and Thomas Icart, "*Password Based Key Exchange Protocols on Elliptic Curves Which Conceal the Public Parameters", ACNS 2010, LNCS 6123*, pp:291-308.

[9] Kakali Chatterjee, Asok De, Daya Gupta, "*Timestamp based Authentication Protocol for Smart Card using ECC", in proceedings of WISM 2011, LNCS 6987*, pp.368-375.

[10] Meng-Hui Lim, Chee-Min Yeoh, Sanggon Lee, Hyotaek Lim and Hoonjae Lee, "*A Secure and Efficient Three-Pass Authenticated Key Agreement Protocol Based on Elliptic Curves" NETWORKING 2008, LNCS 4982*, pp.170-182.

[11] Koblitz, N. 1989, "*Hyperelliptic cryptosystems", Journal of Cryptology 1,3*, pp.139-150.

[12] Adleman L, DeMarrais J,Huang M, "*A subexponential algorithm for discrete. logarithms over the rational subgroup of the Jacobians of large genus hyperelliptic curves over finite fields*", in ANTS-1, 1994, *LNCS 877*, pp.28-40.

[13] Frey G, Ruck H, "*A remark concerning m-divisibility and the discrete logarithm in the divisor class group of curves", Mathematics of Computation*, 1994, 62: pp 865-874.

[14] Ruck H.G "*On the discrete logarithms in the divisor class group of curves". Mathematics Computation*, 1999, 68: 805-806.

[15] Henry Cohen and Gerhard Frey, "*Handbook of Elliptic and Hyperelliptic Curve Cryptography", Chapman & Hall/CRC Press*. 2006.

[16] Menezes A, Wu Y, Zuccherato R, "*An elementary introduction to hyperelliptic curves*", available at http://www.cacr. math.uwaterloo.ca/techreports/1997/tech-reports97.html

[17] J.Pelzl, T.Wollinger, J.Guajardo, C.Paar, "*Hyperelliptic Curve Cryptosystems: Closing the Performance Gap to Elliptic Curves*", *Cryptology ePrint Archieve*, Report 026, http://eprint.iacr.org/, 2003, pp.351-365

[18] Cantor D.G., "*Computing in the Jacobian of a hyperelliptic curve", Mathematics of Computation*, 1987, 48: pp.95-101.

[19] Harly.R, "*Fast Arithmetic on Genus Two Curves*", available at http://cristal. inria.fr/"harly /hyper.

[20] Lange.T, "*Inversion-Free Arithmetic on Genus 2 Hyperelliptic Curves*", *Cryptology ePrint Archieve*, Report 147, 2002, http://eprint.iacr.org/.

[21] Lange.T, "*Efficient Arithmetic on Genus 2 Hyperelliptic Curves over Finite Fields via Explicit Formulae*". *Cryptology ePrint Archive*, Report 121, 2002, http://eprint.iacr.org/.

[22] Roberto Maria Avanzi, "*Aspects of hyperelliptic Curves over Large Prime Fields in Software Implementation*", Dec 2003 available http://www.arehcc.com.

[23] J.Pelzl, T.Wollinger, C.Paar, "*Elliptic & Hyperelliptic Curves on Embedded μP", ACM special issue Security and Embedded Systems* Vol.no.0164-0925/99/0100-0111, 2003.

[24] Gonda.M, Matsuo.K, Kazumaro.A, Chao.J and Tsuji.S, "*Improvements of addition algorithm on genus 3 hyperelliptic curves and their implementations*", *Proc of SCIS* 2004, pp.89-96.

[25] Fan.X, Wollinger.T and Gong.G, "*Efficient explicit formulae for genus 3 hyperelliptic curve cryptosystems over binary fields", IET Inf.Secur.*, 2007,1,(2), pp.65-81.

[26] Kuroki.J, Gonda.M., Matsuo.K., Chao.J., Tsujii. S. 2002, "*Fast Genus Three Hyperelliptic Curve-Cryptosystems", SCIS* 2002, pp.503-507.

[27] Sakai.Y, and Sakurai, K., "*On the Practical Performance of Hyperelliptic Curve Cryptosystems in Software Implementation", in IEICE Trans.* Vol.E83-A NO.4, 2000, pp.692 – 703.

[28] Koh-ichi Nagao, "*Decomposed Attack for the Jacobian of a Hyperelliptic Curve over an Extension Field"*, 2007, http://eprint.iacr.org/ 2007/112.

[29] Benjamin Smith, "*Isogenies and the Discrete Logarithm Problem on Jacobians of Genus 3 Hyperelliptic Curves" EUROCRYPT* 2008, pp.163-180.

[30] Kakali Chatterjee, Daya Gupta, "*Evolution of Hyperelliptic Curve Cryptosystems", in proceedings of ICDCIT 2010, LNCS* 5966, pp.206-211.

[31] Nicholas Jansma, Brandon Arrendondo, "*Performance Comparison of Elliptic Curve and RSA Digital Signatures", University of Michigan*, 2004.

Daya Gupta

She is a Professor and the Head of the Computer Engineering Department at Delhi Technological University in India. She has received her Ph.D. in Computer Engineering from Delhi University. Her field of interest is Software Engineering, Information Security, etc. She has published many research papers in reputed international journals.

Asok De

He received his Ph.D. from IIT Kharagpur (India) and his field of interest is Microwave Antennas and Communication Systems. He is a Professor at the Delhi Technological University (formerly the Delhi College of Engineering). Presently he is working as the Principal at the Ambedkar Institute of Advanced Communications Technologies & Research in Delhi. He has published many research papers in reputed international journals.

Kakali Chatterjee

She is a Research Scholar in the Computer Engineering Department of Delhi Technological University (formerly the Delhi College of Engineering) in India. She has received her M.Tech (Information Technology) from the Centre for the Development of Advanced Computing, which is a R&D and academic centre of the govt. of India. Her field of interest is Information Security and Cryptography.

Social Networking based E-Learning System on Clouds

Rajni Jindal
Associate Professor
Department of Computer Engineering,
Delhi Technological University,New Delhi

Alka Singhal
Research Scholar
Department of Computer Engineering,
Delhi Technological University,New Delhi

ABSTRACT

With the recent advancements in the modern Information and Communication Technology (ICT), e-Learning has emerged as a new paradigm for learning in the modern world. There are many dimensions such as pedagogical, technological, ethical etc which are to be satisfied by the e-learning service provider to become a better option in compare to the traditional learning techniques. Among all the dimensions, technological and pedagogical dimension are among the critical dimensions, as they address issues concerning content analysis, audience analysis, goal analysis, performance analysis and infrastructural analysis. This paper proposes an E-learning Social networking site which is maintained by Cloud providers. Blending the two technologies, Social networking and Cloud computing, provides a business model for E-learning where construction of e-learning system is entrusted to cloud computing suppliers and social networking helps to improve the teaching quality and content.

Keywords

E-learning Systems, Social networking, Cloud computing, Social clouds.

1. INTRODUCTION

E-learning as an important mode of learning today, plays an important role in creating a good convenient learning environment. It is a good carrier of the content and provides learner a wide variety of learning materials and learning opportunities. It has lots of advantages like flexibility, diversity, measurement etc and so it is becoming a primer way for learning in the new century [1].

Key factors of E-learning are:-1) Reuse 2) Resource Sharing 3) Interoperability.

E-learning can be viewed as a way of learning in which the instructor and the student are separated by distance or time and this gap is bridged through the use of online resources. It can be any of the following technology:-

- Blogs.
- Computer aided teaching.
- Online Discussion Boards.
- Electronic education support system.
- Learning management systems.
- Virtual classrooms.
- Web-based teaching materials.

First part of the paper includes Infrastructural analysis of the E-learning system including content distribution and expansion. In traditional web-based e-learning mode, system construction and maintenance are done by the educational institutions itself. As it involves large investment and maintenance from the institution, the capital gain is not much as it should be. This is also hindering the E-learning development in the computing era. In contrast, if cloud computing is used in the e-learning model, it introduces scalability, i.e. construction and maintenance of e-learning system is done by the cloud services providers and further these services are used by the e-learning provider by paying in cost per unit. Here E-learning provider is not concerned about the infrastructural and maintenance expenses but it focus on the content quality and management [2]. Whereas, the Cloud providers have already established infrastructure, so they can utilize it for providing E-learning services without any additional investment in compare if it is done by Institution itself. In totality, it gives a Win-Win situation to both and also provides better prospects for E-learning systems.

Second part of the paper explores the universality and pedagogical dimension of the e-learning content. There is a huge amount of data available on the web, but this data is many times not accessible due to its different language and locality. The social cloud helps all the e-learning providers a common platform where they can share their data and make it universally available. The model will act as knowledge search engine where all the contents of providers will be available to the learner. This search engine will act as a social networking site between all the providers and learners. Social Clouds help learners to interact with each other and perform discussions. Learners can also give their feedbacks for the material and thus content will be well reviewed and updated. In summarized way, this paper proposes a social networking e-learning site which is maintained by the cloud providers.

2. TRADITIONAL E-LEARNING MODEL

Traditional e-learning network is located in a campus network or an Intranet with its construction, maintenance, and investment being made by enterprise itself. There are various components of E-learning System [3]:-

Content Experts:-Content Experts are responsible for production of authorized learning material. Material can be in any form e.g. lectures notes, video courses etc. They are

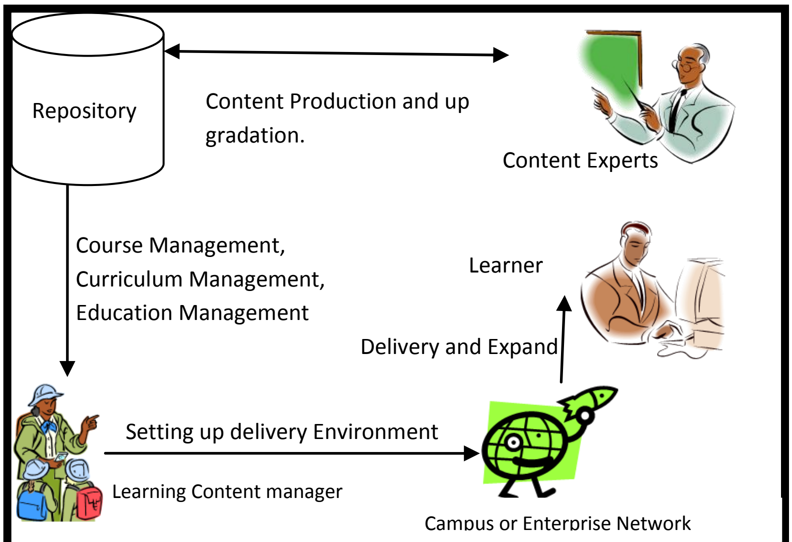


Fig1.Traditional E-learning system

esponsible for creating good, reliable and updated learning material for the targeted learner.

Metadata & Repository: - It contains all the learning content in form of lectures, videos etc and also the catalogs maintaining the information about the learning content like indexes, date of issue, relevant topics, keywords etc.

Learning Content Manager: A manager who is the administrator of the E-learning system, he keeps information about the registered learners, their payments, their feedbacks. He is responsible for structuring data, organizing reviews, tests etc.

Content Distribution:-Content distribution system is responsible for utilizing wide area network, broadband services to distribute the data. It is responsible for error-free and fast delivery of data. As it can be real-time system (virtual classrooms) time complexity is important.

Learner:-Learner is the one accessing the E-learning system. E-learning system should maintain the information about learner profile and his registration information. Learner profile information can include personal data, learning plans, learning history, accessibility requirements, certifications and degrees, assessments of knowledge and the status of participation in current learning.

3. USING SOCIAL NETWORKING IN E-LEARNING

A social network is a set of individuals connected through socially meaningful relationships, such as friendship or information exchange (Wasserman, Faust, Iacobucci, & Granovetter, 1994; Wellman, 1996).

Social networks are formed when people interact with each other (Garton, Haythornthwaite, & Wellman, 1997) and thus can be seen in many aspects of everyday life. In its most simple form, a social network is a map of all of the relevant links between the nodes being studied. As noted by Milgram (1967), the strength of a tie between two actors is much greater if they have another mutual acquaintance. In other words, the probability of two friends of an individual knowing each another is much greater than the probability of two people chosen randomly from the population knowing each another (Guare, 1990; Newman, Watts, & Strogatz, 2002; Watts & Strogatz, 1998). Actors with strong ties usually have some sort of common ground on which they establish their relationships (Preece, 2002; Wellman & Gulia, 1997), and thus often constitute a subgroup. Because of the common ground, actors with strong ties – and hence representing a subgroup – often share common interests, needs, or services that provide a reason for the subgroup (Preece, 2002; Schwartz & Wood, 1993; Wellman & Gulia, 1997).

A Social networking can be widely explored in the field of E-learning as it is said "Knowledge which is shared is better". So the paper provides a platform where learners of common interest as well as providers who are commonly targeted by the learners are grouped together to form a network. This platform will help the various e-learning providers to communicate and share their contents and spread it to a larger audience. On the other side, learners having common interest will share and communicate their views etc. As a whole it will give a new paradigm for E-earning systems.

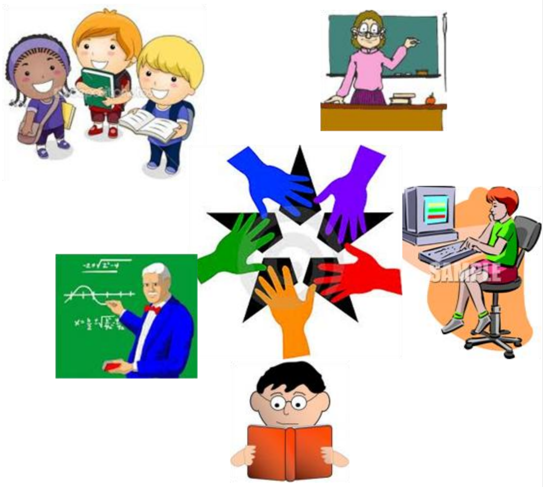


Fig2. Social Networking in E-Learning

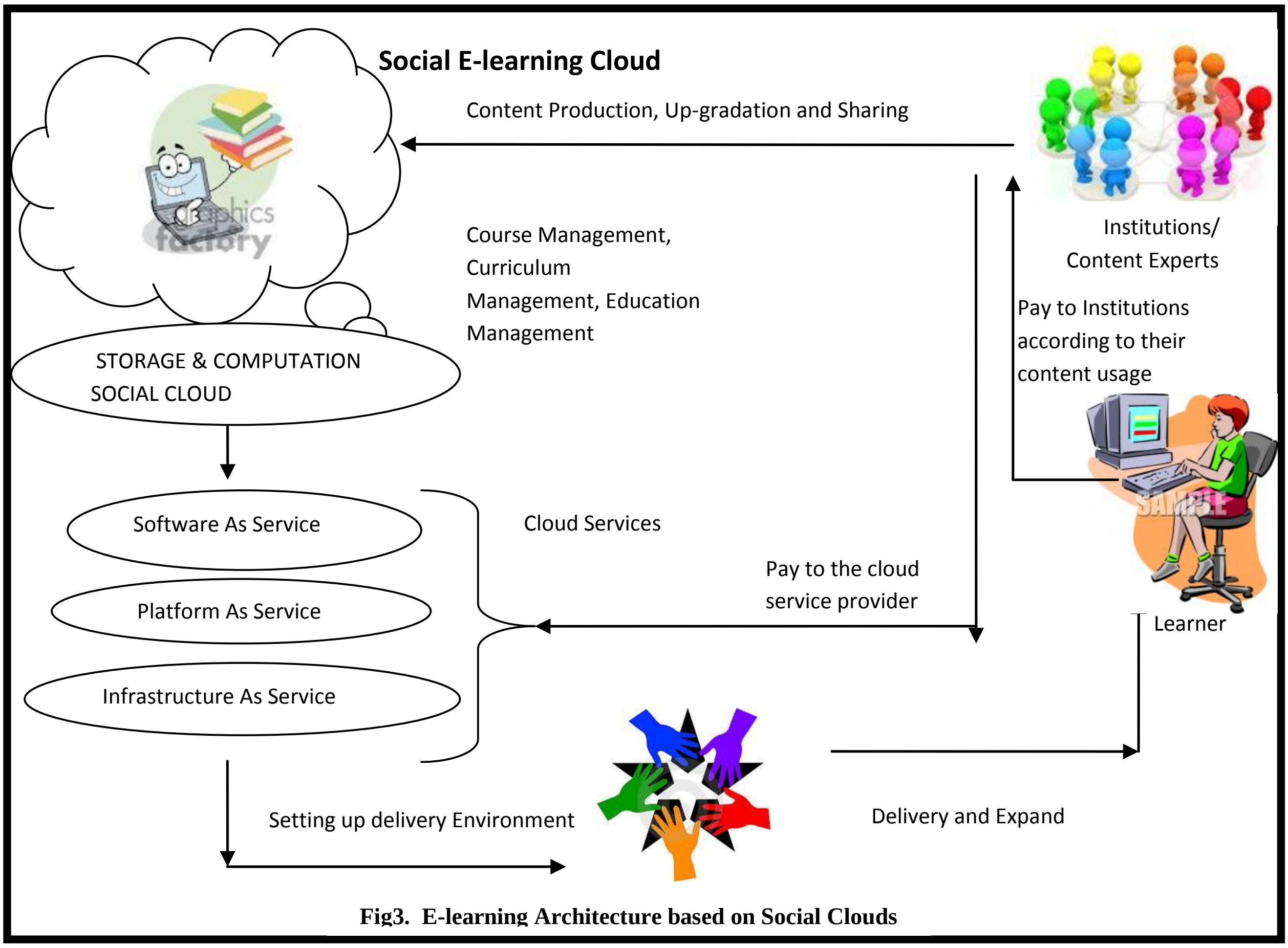

Social E-learning Cloud
Content Production, Up-gradation and Sharing
Course Management, Curriculum Management, Education Management
Institutions/ Content Experts
Pay to Institutions according to their content usage
STORAGE & COMPUTATION SOCIAL CLOUD
Software As Service
Platform As Service
Infrastructure As Service
Cloud Services
Pay to the cloud service provider
Learner
Setting up delivery Environment
Delivery and Expand

Fig3. E-learning Architecture based on Social Clouds

4. INFRASTRUCTURAL SUPPORT FROM CLOUDS

Today is the age of information technology. The facets of work and personal life are moving towards the concept of availability of everything online. In e-learning cloud computing business model, cloud provider is responsible for building and maintaining e-learning cloud, providing technical support to the e-learning cloud. Cloud users pay to the cloud provider for services and services are accessed on-demand. E-learning cloud is a migration of cloud computing technology in the field of e-learning, which is a future e-learning infrastructure, including all the necessary hardware and software computing resources engaging in e-learning [8]. After these computing resources are virtualized, they can be afforded in the form of services for educational institutions, students and businesses to rent computing resources [9][10].At the lowest level, the institutions which will provide the e-learning service produce courses, content management and its up-gradation and maintenance. E-learning Cloud provides suitable resources for storage and computation. Then the service layers in cloud computing, namely, SaaS (Software as a Service), PaaS (Platform as a Service), IaaS (Infrastructure as a Service) provides services to cloud users (Institutions) for Education management, Course management etc. On the other hand it provides environment for delivery to the learner. Learner can be at any remote place. Cloud services help in efficient delivery. Learner pays to the Institution for courses undertaken and in response cloud users (Institutions) pay Cloud providers for their services.

5. E-LEARNING ON SOCIAL CLOUD

With the increasingly usage of Social networks and Cloud computing, users are starting to explore new ways to interact with, and exploit these developing paradigms. Social networks are the relationships that allow users to share information and forming Virtual Organizations. We propose e-learning model which utilize institutions having pre trust in each other to form a “Social Cloud”, enabling institutions to share resources within the context of a Social network.

There are already existing Social networks such as MyExperiment and NanoHub for research community. They provide a virtual research environment where collaborators can share research and execute scientific workflows remotely. The same functionality can be realized using a Social Cloud deployed in an existing Social network. For example Social Storage Clouds can be used to store/share data and information (for example academic papers, scientific workflows, datasets, and analysis) within a community [7].

According to user’s preference, he/she can find and combine useful learning material conveniently and economically. Each course can be assigned credentials by the respective college and required payments are done by the learner, according to the usage. At the front there is a social networking site where all the colleges are registered. Learners are assigned with a unique User id for authentication. After login, they are given a wide variety of courses offered by various providers registered. According to one’s preference, a set of courses are offered. Learner is charged according to the courses. Each user's course information is stored in the Storage Cloud. Payment is redeemed in respective college accounts using Cloud services. Infrastructural support and other services are given by Cloud service provider as specified above. Cloud provider's payment is shared by the social network registered colleges and so providers (institutions) can be free from the building and maintenance for e-learning system and specifically focus on the application of e-learning system in order to improve teaching quality and management level.

6. ADVANTAGES AND CHALLENGES

The proposed model has following advantages:-

1. The service will be faster, simpler and on demand.
2. The E-learning system resources will be more scalable related to overcrowded and lean periods. There will be optimized use of computing resources. Cloud will easily handle peak load situation without any additional infrastructural support.
3. As everything will be a service (SaaS, PaaS, IaaS), Learners will be having unlimited storage and resources which will give them flexibility to explore their ideas.
4. There will be high availability and reliability. As data is stored in various data centers of the cloud providers, there is less chance of data loss. Learners can use any electronic device connected with internet to access their E-learning system.
5. As Cloud computing is bringing new technologies in computing era to reduce cost and increase computing efficiency, it will be indirectly reflected in the E-learning Era.
6. Cloud computing infrastructure allows enterprises to achieve more efficient use of their IT hardware and software investments: it increases profitability by improving resource utilization. Pooling resources into large clouds cuts costs and increases utilization by delivering resources only for as long as those resources are needed[11].
7. Cloud computing infrastructure can be located in areas with lower costs of space and electricity.

The propose model still has many following challenges to become a practical working model [11]:-

1. To obtain an accurate cost and charging model :- From a cloud provider's perspective, the elastic resource pool (through either virtualization or multi-tenancy) has made the cost analysis a lot more complicated than regular data centers, which often calculates their cost based on consumptions of static computing.
2. Security: - Security is a major issue for the cloud computing as well as social networking. Using a third party to keep and manage your data, requires faith in the cloud provider and also while sharing data in a social networking site; the authentication of the original work should be maintained by the cloud.
3. Service Level Agreement: - Cloud consumers need to ensure the quality, availability, reliability, and performance of the service provides. In other words, it is necessary for consumers to obtain guarantees from providers on service delivery. Typically, these are provided through Service Level Agreements (SLAs) negotiated between the providers and consumers.

7. CONCLUSION AND FUTURE WORK

This paper has presented the architecture and implementation of a Social Cloud with E-learning; a combination of Cloud computing, Social networking and E-learning. Exploring the benefits of all technologies, it provides a business model where construction and maintenance is done by cloud providers and all the E-learning providers merge to form a social group to share and enrich the teaching content as well as teaching quality. Though there are lots of challenges like charging market mechanism, Bandwidth for the content distribution, Security etc which are in front of success of the model.

8. REFERENCES

[1] Liang Bing, "E-learning and modern education reform", Education Information, 2001.10, pp.21, 25

[2] Zhu Chengyun, "Cloud Security: The security risks of cloud computing, models and strategies", Programmer, May.2010, pp.71- 73

[3] Xiaofei Liu, Abdulmotaleb El Saddik and Nicolas D. Georganas,, AN IMPLEMENTABLE ARCHITECTURE OF AN E-LEARNING SYSTEM,CCECE 2003, Montreal, May 2003

[4] Wang Xiaomei, Jia Xiaoqiang, "Cloud computing on the Impact of Higher Education", Science & Technology Information, 2010.10, pp.397-398

[5] ZHao Zhong- ping, LIU Hui-cheng , "The Development and Exploring of E- Learning System on Campus Network", Journal of Shanxi Teacher's University (Natural Science Edition), Vol.18, No.1, Mar. 2004, pp.36-40

[6] Xu Chuanling, Lu Hongjie, "E-learning", Software World, 2001.08, pp. 139-141

[7] Kyle Chard, Simon Caton, Omer Rana, Kris Bubendorfer, "Social Cloud: Cloud Computing in Social Networks" 2010 IEEE 3rd International Conference on Cloud Computing.

[8] Xiao Laisheng,Wang Zhengxia, "Cloud Computing: a New Business Paradigm for E-learning", 2011 Third International Conference on Measuring Technology and Mechatronics Automation

[9] Liu Huanying, "Value and understanding for cloud computing based on middleware" ,Programmer, 2010.05. pp.68,69

[10] Fu feng, "Cloud-based IT infrastructure of next-generation telecom", Mobile Communications, 2010, No. 8, pp.76-79

[11] Tharam Dillon, Chen Wu and Elizabeth Chang," Cloud Computing: Issues and Challenges" 2010 24th IEEE International Conference on Advanced Information Networking and Applications.

The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on Decision Tree

Nishant Mathur, Sumit Kumar, Santosh Kumar, and Rajni Jindal

Abstract—**Data mining is used to extract the required data from large databases [1]. The data mining algorithm is the mechanism that creates mining models [2]. To create a model, an algorithm first learns the rules from a set of data then looks for specific required patterns and trends according to those rules. The algorithm then uses the fallouts of this exploration to delineate the constraints of the mining model [2]. These constraints are then applied through the intact data set to extract the unlawful patterns and detailed statistics [2]. Decision-tree learning is one of the utmost efficacious erudition algorithms, due to its various eye-catching features: simplicity, comprehensibility, no parameters, and being able to handle mixed-type data [3]; ID3 is a simple decision tree erudition algorithm developed by Ross Quinlan (1983) [4]. This paper introduces the use of ID3 algorithm [4] of decision tree and we use Havrda and Charvat Entropy instead of Shannon Entropy [5]. By computing information we set particular property from taken data as root of tree, also sub-root by repeating the process continually, to finally build the most optimized tree. This decision tree helps to take the decision for better analysis of data. Decision tree algorithm is used to select the best path to follow in the standard division. This paper introduces the use of ID3 algorithm of decision tree. We are using Havrda and Charvat Entropy Instead of Shannon Entropy. This Decision Tree helps in taking the better decision to analyse the data.**

Index Terms—**Data mining, decision tree, Shanon entropy, Havrda and Charvat entropy, ID3 algorithm, knowledge-driven decisions.**

I. Introduction

Data mining is the technique to extract the hidden predictive data from large databases; it is an influential technology and used by lot of companies because of very prodigious fallouts [6], [7]. Data mining is very supportive technique to analyse the forthcoming prediction with the help of historical behaviour of data and statistics, these features of data mining sanction proactive business and it is called knowledge-driven decisions. This automation, prospective scrutinizing and exploration of past events work as retrospective tool and implement a DSS (decision support system).

Data mining techniques can rapidly implement on existing software and hardware and intensify the quality of service of them.

A. ID3 Algorithm

ID3 is a simple decision tree erudition algorithm developed by Ross Quinlan (1983) [4]. The basic idea of ID3 algorithm is to create a decision tree of given set, by using top-down greedy search to check each attribute at every tree node. To select the most useful attribute using classification technique, we present a metric---information gain and to catch an optimal way to classify an erudite set, we need to minimize the depth of the tree. Thus, we need some function which should be able to measure the most balanced splitting. The information gain metric is such a function that we can use for efficient balanced splitting. In direction to define information gain exactly, we need to deliberate entropy. First, let's assume that the resulting decision tree classifies instance into two classes without loss of simplification and we would call them P (positive) and N (negative).

Given set S, containing these positive and negative targets, the entropy of S related to this Boolean classification is:

Entropy(S) = −P (positive) log2P (positive) − P (negative) log2P (negative)

P (positive): proportion of positive examples in S
P (negative): proportion of negative examples in S

So concerning Points are as we discussed, to minimize the decision tree depth; we need to select the optimal attribute for splitting the tree node, so that we can easily imply the attribute with the maximum entropy reduction. The attribute that can help in maximum entropy reduction is the optimal attribute for splitting. We define Information Gain as the predictable reduction of entropy related to specified attribute when splitting a decision tree node. The information gain, Gain(S, A) of an attribute A,

Gain(S, A) = Entropy(S) −Sum for v from 1 to n of (|Sv| ⁄|S|) x Entropy (Sv).

We have to use this concept of gain to rank attributes to build decision trees where at each node is located the attribute with utmost gain among the attributes that not yet considered in the path from the root.

The purpose of this ordering is to create small decision trees so that records can be identified after only a few

Manuscript received January 28, 2012; revised March 3, 2012.

N. Mathur is with the Delhi College of Engineering, India (e-mail: javadce@gmail.com).

Sumit Kumar and Santosh Kumar are with the Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India (e-mail: sumit.itech@gmail.com, mrsonuk@gmail.com).

R. Jindal is with Delhi Technological University, India (e-mail: rajnijindal@dce.ac.in).

decision tree splitting and match a hoped for plainness of the process of decision making.

B. Problem Statement

Shannon Entropy finds its application in many fields. Here, Shannon Entropy has been used in ID3 algorithm to calculate the Information Gain contained by data, which helps to make Decision Tree.

However, the results obtained from Shannon Entropy, are rather complex, have more numbers of node and leaf and Decision Rules. Thus it makes the decision making process time consuming.

Therefore, to minimize these problems, new algorithm has been proposed by modifying ID3 algorithm using Havrda and Charvat Entropy instead of Shannon Entropy

II. Evaluation and Depiction

Classification is perhaps the utmost acquainted and the most extensive data mining technique. Examples of classification application are images and pattern recognition, medical diagnosis, loan approval, detecting faults in industrial application, and classifying market trends [8]. Estimation and prediction can be view as types of classification. Prediction can classify an attribute value from a set of possible values. It is often viewed as forecasting value, while classification forecasts a discrete value.

All methodologies to perform classification assume certain acquaintance of the data. Habitually a training set is used to develop the precise parameters, those are obligatory through the technique. Training data contains a sample of input data as well as the classification assignment for the data.

The classification problem is stated as:

Definition: Given a database $D = \{t_1, t_2, t_3 \ldots\ldots t_m\}$ of tuples (items, records) and a set of classe $C = \{c_1, c_2, c_3 \ldots.. c_m\}$ the classification problem is to define a mapping f: D-> C where t_i is assigned to one class. A class, c_j contain precisely those mapped to it; that is, $c_j = \{t_i / f(t_i) = c_j$, $1 <= i <= n$, and t_1 belong to $D\}$

According to definition's interpretations, classification is a mapping of the database to the set of classes. Each tuple in the database is assigned to exactly one class.

The classes that exist for a classification problem are indeed equivalence classes. In actuality, the problem usually is implemented in two phases:

1) Create a specific model by evaluating the training data. This step takes the training data as input and gives the output as the definition of the developed model. The developed model classifies the training data as accurate as possible.
2) Apply the established model in step 1 by classifying tuples from the target database.

A. Depiction of Decision Tree and ID3 Algorithm

In this section, we define the ID3 Algorithm and the depict role of decision tree. The Decision Trees algorithm creates hierarchical structure of classification rules "If ... Then ..." looking like a tree. To choose which type to assign for an object or state, we have to answer the questions, standing in the branches of the tree, starting from the root. The questions look like this: "Is the value of the parameter A greater than X?" If the answer is positive, a pass to the right performs, if it is negative – to the left; then a question related to the new branch follows. In following illustration, information about customers was produced, including their debt level, income level, what type of occupation they had, and whether they represented a good or bad credit risk

A decision Tree consists of 3 types of nodes:

1) Decision nodes - commonly represented by squares
2) Change nodes - represented by circles
3) End nodes - represented by triangles

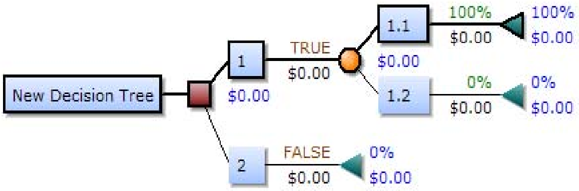


Fig. 1. Decision tree showing types of nodes.

Decision trees are used in operations research, decision analysis, to help identify a strategy, calculating conditional probabilities etc.

In decision analysis, a "decision tree", is used as a visual and analytical decision support tool, where the predictable value of competing options are deliberate.

Decision trees have traditionally been created manually, as shown in Fig. 1.

III. Issues, Implementation and Results of Proposed Algorithm using Havrda and Charvat Entropy

The ID3 algorithm is used to build a decision tree, given a set of non-categorical attributes C1, C2, .., Cn, the categorical attribute C, and a training set T of records.

Algorithm

1) Function ID3 (R: a set of non-categorical attributes,
2) C: the categorical attribute,
3) S: a training set) returns a decision tree;
4) begin
5) If S is empty, return a single node with value Failure;
6) If S consists of records all with the same value for the categorical attribute,
7) Return a single node with that value;
8) If R is empty, then return a single node with as value the most frequent of the values of the categorical attribute that are found in records of S; [note that then there will be errors, that is, records that will be improperly classified];
9) Let D be the attribute with largest Gain (D,S)
10) Among attributes in R;
11) Let $\{dj | j = 1, 2, .., m\}$ be the values of attribute D;
12) Let $\{Sj | j = 1, 2, .., m\}$ be the subsets of S consisting respectively of records with value dj for attribute D;
13) Return a tree with root labeled D and arcs labelled.

```
14) $d_1, d_2, .., d_m$ going respectively to the trees
15) ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), .., ID3(R-{D}, C, Sm);
16) end ID3;
```

A. *Entropy as Information Content*

Entropy is demarcated in the perspective of a probabilistic model. Independent fair penny tosses have Entropy of 1 bit per flip. A source that always produces a long string of B's has Entropy of 0, since the next character will always be a 'B'. The entropy rate of a data source means the average number of bits per sign needed to encode it.

B. *Definition and Role of Havrda and Charvat Entropy*

Let P = $(p_1, p_2 \ldots p_n)$ be a probability distribution, p denotes the probability mass function of X and α is its inherent parameter [13]. Then Havrda and Charvat [12] gave the entropy measure by formula shown under

$$h(p) = \frac{1}{1-a}\left(\sum_{i=1}^{n} X_i^a - 1\right)$$

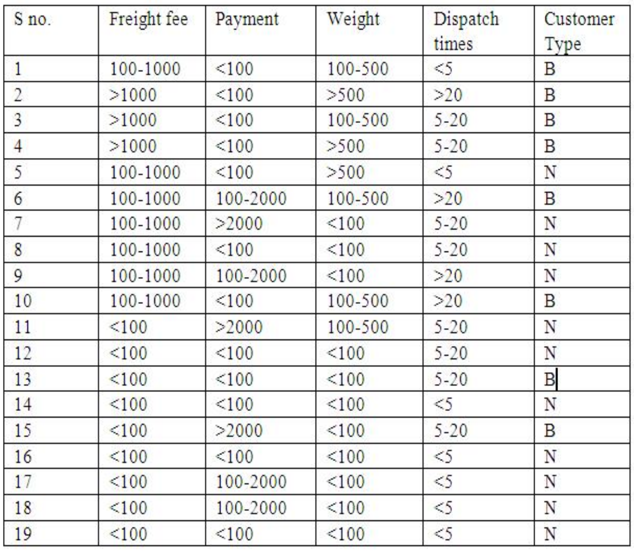
This formula calculates Entropy. To avoid deduced solution in decision tree making process, Havrda and Charvat entropy based ID3 algorithm is proposed which gives good solution in reasonable time. Such algorithm can give short and fast decision for supply of good in company.

C. *Havrda and Charvat Entropy in ID3 algorithm*

The measure of tree component is one of the most important problems of ID3. Such problems occur when we have to take decision for who will be the root of the tree. So to find one we have to calculate first the needed information, so for needed information we divided our data [2] into category for which we are making tree here we divided data according to customer type, here taking sample data into two parts: Sing_Customer (B) and Normal_Customer (N).

D. *Implementation and Analysis of Proposed Algorithm*

TABLE I: INFORMATION OF CUSTOMER DISPATCH GOODS



S no.	Freight fee	Payment	Weight	Dispatch times	Customer Type
1	100-1000	<100	100-500	<5	B
2	>1000	<100	>500	>20	B
3	>1000	<100	100-500	5-20	B
4	>1000	<100	>500	5-20	B
5	100-1000	<100	>500	<5	N
6	100-1000	100-2000	100-500	>20	B
7	100-1000	>2000	<100	5-20	N
8	100-1000	<100	<100	5-20	N
9	100-1000	100-2000	<100	>20	N
10	100-1000	<100	100-500	>20	B
11	<100	>2000	100-500	5-20	N
12	<100	<100	<100	5-20	N
13	<100	<100	<100	5-20	B
14	<100	<100	<100	<5	N
15	<100	>2000	<100	5-20	B
16	<100	<100	<100	<5	N
17	<100	100-2000	<100	<5	N
18	<100	100-2000	<100	<5	N
19	<100	<100	<100	<5	N

The summarized data of customer dispatch information in a section period (one month) from an information system database of a 3PL, which including 19 items in this sample data set. In this example [13], all sample data is divided by Customer Type (CT) into two classes, which are Sign_Customer (B) and Normal_Customer (N) respectively, and has four properties: Freight Fee, Payment, Weight, and Dispatch Time. On the one hand, the summarizing data is integrated data from different sections and different consignment nodes. On the other hand, it is the process of generalizing the sample data, namely, the low level data are substituted by high level convenient to data mining. The values of these four properties are: Freight Fee (<100,100~1000, >1000); Payment (<100, 100~2000, <2000); Weight (<100 kg, 100 kg~500 kg, >500 kg); Dispatch Times (<5, 5~20, >20) [2]. The meanings of these properties are: The freight fee is paid by customer for the transport cost; the payment is Transportation Company bring the money of the goods from the receiver to dispatcher; the weight is measured by kilogram; the dispatch time is the sum times during the summarized period.

We can calculate needed information by taking probability of customer type here we B class have 8 items and N has 11 items. Therefore, needed information gain of taken sample by putting α=0.25, 0.50, 0.75 etc in Havrda and Charvat formula.

Assuming α=0.25

The needed information gain will be

$$I(8,11) = \frac{\left[\left(\frac{8}{19}\right)^{.25} + \left(\frac{11}{19}\right)^{0.25} - 1\right]}{1-0.25} = \frac{0.677}{0.75} = 0.903$$

Then we divided the sample data by four property, freight fee, payment, weight, and dispatch time respectively. Therefore the corresponding anticipated information of sample data are:

$$\text{E (Freight fee)} = \frac{\frac{9}{19}\left[\left(\frac{2}{9}\right)^{.25} + \left(\frac{7}{9}\right)^{0.25} - 1\right]}{0.75} + \frac{\frac{7}{19}\left[\left(\frac{4}{7}\right)^{.25} + \left(\frac{3}{7}\right)^{0.25} - 1\right]}{0.75} + \frac{\frac{3}{19}\left[\left(\frac{3}{3}\right)^{.25} - 1\right]}{0.75} = 0.395 + 0.333 + 0 = 0.728$$

$$\text{E (Payment)} = \frac{\frac{12}{19}\left[\left(\frac{6}{12}\right)^{.25} + \left(\frac{6}{12}\right)^{0.25} - 1\right]}{0.75} + \frac{\frac{4}{19}\left[\left(\frac{1}{4}\right)^{.25} + \left(\frac{3}{4}\right)^{0.25} - 1\right]}{0.75} + \frac{\frac{3}{19}\left[\left(\frac{2}{3}\right)^{.25} + \left(\frac{1}{3}\right)^{0.25} - 1\right]}{0.75} = 0.574 + 0.179 + 0,139 = 0.893$$

$$\text{E (Weight)} = \frac{\frac{11}{19}\left[\left(\frac{2}{11}\right)^{.25} + \left(\frac{9}{11}\right)^{0.25} - 1\right]}{0.75} + \frac{\frac{5}{19}\left[\left(\frac{4}{5}\right)^{.25} + \left(\frac{1}{5}\right)^{0.25} - 1\right]}{0.75} + \frac{\frac{3}{19}\left[\left(\frac{2}{3}\right)^{.25} + \left(\frac{1}{3}\right)^{0.25} - 1\right]}{0.75} = 0.466 + 0.216 + 0.139 = 0.822$$

$$E\ (\text{Time}) = \frac{\frac{7}{19}\left[\left(\frac{1}{7}\right)^{.25}+\left(\frac{6}{7}\right)^{0.25}-1\right]}{0.75} + \frac{\frac{8}{19}\left[\left(\frac{4}{8}\right)^{.25}+\left(\frac{4}{8}\right)^{0.25}-1\right]}{0.75} + \frac{\frac{4}{19}\left[\left(\frac{3}{4}\right)^{.25}+\left(\frac{1}{4}\right)^{0.25}-1\right]}{0.75} = 0.283 + 0.383 + 0.179 = 0.845$$

Corresponding information gains are:

Gain (Freight fee) = I (S1, S2) – E (Freight fee)

= 0.903 – 0.728 = 0.175

Gain (Payment) = I (S1, S2) – E (Payment)

= 0.903 – 0.893 = 0.011

Gain (Weight) = I (S1, S2) – E (Weight)

= 0.903 – 0.822 = 0.082

Gain (Time) = I (S1, S2) – E (Time)

= 0.903 – 0.845 = 0.058

The information gain for freight fee is largest. Therefore freight fee will be root of decision tree. Now considering Table 2 where freight fee is <100, from main table.

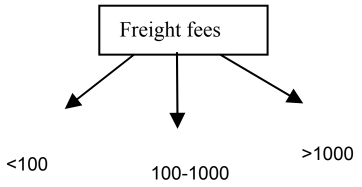


Fig. 2. Root of decision Tree

TABLE II: INFORMATION OF CUSTOMER DISPATCH GOODS FOR FREIGHT FEE (<100)

S no.	Freight fee	Payment	Weight	Dispatch times	Customer Type
11	<100	>2000	100-500	5-20	N
12	<100	<100	<100	5-20	N
13	<100	<100	<100	5-20	B
14	<100	<100	<100	<5	N
15	<100	>2000	<100	5-20	B
16	<100	<100	<100	<5	N
17	<100	100-2000	<100	<5	N
18	<100	100-2000	<100	<5	N
19	<100	<100	<100	<5	N

So, here B=2, N= 7 we calculated needed information

$$I\ (2,\ 7) = \frac{\left[\left(\frac{2}{9}\right)^{.25}+\left(\frac{7}{9}\right)^{0.25}-1\right]}{0.75} = 0.625/.75 = 0.834$$

Corresponding anticipated information of different properties is:

E (Freight fee) = 0.834

$$E\ (\text{Payment}) = \frac{\frac{5}{9}\left[\left(\frac{1}{5}\right)^{.25}+\left(\frac{4}{5}\right)^{0.25}-1\right]}{0.75} + \frac{\frac{2}{9}\left[0+\left(\frac{2}{2}\right)^{0.25}-1\right]}{0.75} + \frac{\frac{2}{9}\left[\left(\frac{1}{2}\right)^{.25}+\left(\frac{1}{2}\right)^{0.25}-1\right]}{0.75} = 0.455 + 0 + 0.202 = 0.657$$

$$E\ (\text{Weight}) = \frac{\frac{8}{9}\left[\left(\frac{2}{8}\right)^{.25}+\left(\frac{6}{8}\right)^{0.25}-1\right]}{0.75} + 0 + 0 = 0.756$$

$$E\ (\text{Time}) = \frac{\frac{5}{9}\left[0+\left(\frac{5}{5}\right)^{0.25}-1\right]}{0.75} + \frac{\frac{4}{19}\left[\left(\frac{2}{4}\right)^{.25}+\left(\frac{2}{4}\right)^{0.25}-1\right]}{0.75} + 0$$

= 0 + 0.404 + 0 = 0.404

Now the corresponding Information Gain for above properties is:

Gain (Freight fee) = I (S1, S2) – E (Freight fee)

= 0.834 – 0.834 = 0

Gain (Payment) = I (S1, S2) – E (Payment)

= 0.834 – 0.657 = 0.177

Gain (Weight) = I (S1, S2) – E (Weight)

= 0.834 – 0.756 = 0.078

Gain (Time) = I (S1, S2) – E (Time)

= 0.834 – 0.404 = 0.430

Here Information gain of Time is maximum, therefore Time will be SUBROOT under Freight Fee (<100) root.
Now consider Table II for freight fee (100-1000) from main table.

TABLE III: INFORMATION OF CUSTOMER DISPATCH GOODS FOR FREIGHT FEE (100-1000)

S no.	Freight fee	Payment	Weight	Dispatch times	Customer Type
1	100-1000	<100	100-500	<5	B
5	100-1000	<100	>500	<5	N
6	100-1000	100-2000	100-500	>20	B
7	100-1000	>2000	<100	5-20	N
8	100-1000	<100	<100	5-20	N
9	100-1000	100-2000	<100	>20	N
10	100-1000	<100	100-500	>20	B

Here B=3 and N=4

Now the Needed Information Gain here will be,

$$I\ (3,\ 4) = \frac{\left[\left(\frac{3}{7}\right)^{.25}+\left(\frac{4}{7}\right)^{0.25}-1\right]}{0.75} = 0.905$$

Also, corresponding anticipated information for different properties are:

E (Freight Fee) = 0.905

$$\text{E (Payment)} = \frac{\frac{4}{7}\left[\left(\frac{2}{4}\right)^{.25}+\left(\frac{2}{4}\right)^{0.25}-1\right]}{0.75} + \frac{\frac{2}{7}\left[\left(\frac{1}{2}\right)^{.25}+\left(\frac{1}{2}\right)^{0.25}-1\right]}{0.75} + \frac{\frac{1}{7}\left[0+\left(\frac{1}{1}\right)^{0.25}-1\right]}{0.75} = 0.519 + 0.259 + 0 = 0.778$$

$$\text{E (Weight)} = \frac{\frac{3}{7}\left[0+\left(\frac{3}{3}\right)^{0.25}-1\right]}{0.75} + \frac{\frac{3}{7}\left[0+\left(\frac{3}{3}\right)^{0.25}-1\right]}{0.75} + \frac{\frac{1}{7}\left[\left(\frac{1}{1}\right)^{0.25}-1\right]}{0.75} = 0 + 0 + 0 = 0$$

$$\text{E (Time)} = \frac{\frac{2}{7}\left[\left(\frac{1}{2}\right)^{.25}+\left(\frac{1}{2}\right)^{0.25}-1\right]}{0.75} + \frac{\frac{3}{7}\left[\left(\frac{3}{3}\right)^{.25}-1\right]}{0.75} + \frac{\frac{2}{7}\left[\left(\frac{2}{2}\right)^{.25}-1\right]}{0.75} = 0.638 + 0 + 0 = 0.638$$

The corresponding information gains are:

Gain (Freight fee) = I (S1, S2) – E (Freight fee)

= 0

Gain (Payment) = I (S1, S2) – E (Payment)

= 0.905 – 0.778 = 0.127

Gain (Weight) = I (S1, S2) – E (Weight)

= 0.905 – 0 = 0.905

Gain (Time) = I (S1, S2) – E (Time)

= 0.905 – 0.638 = 0.265

Information Gain for Weight is largest, so it will be sub root, under Freight fee (100-1000) category.

Now Consider table for Freight fee (>100)

TABLE IV: INFORMATION OF CUSTOMER DISPATCH GOODS FOR FREIGHT FEE (>1000)

S no.	Freight fee	Payment	Weight	Dispatch times	Customer Type
2	>1000	<100	>500	>20	B
3	>1000	<100	100-500	5-20	B
4	>1000	<100	>500	5-20	B

From Table IV we can conclude that under freight fee (100-1000) sub root will be PAYMENT where for it <100 customers will be B type only. Same way we will conclude for sub root weight under (freight fee 100-1000 only) from table we can observe that for sub root weight <100 customer is N type, for 100-1000 customer is B type and for weight >500 customer is N type. From above calculation and observation we have drawn the following tree.

E. Generation of Decision Tree

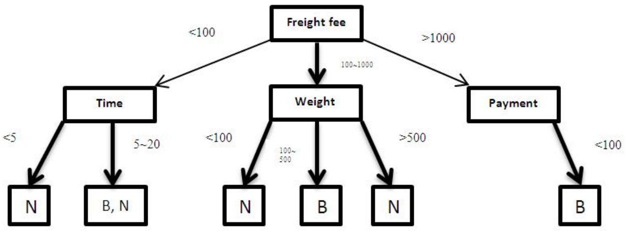
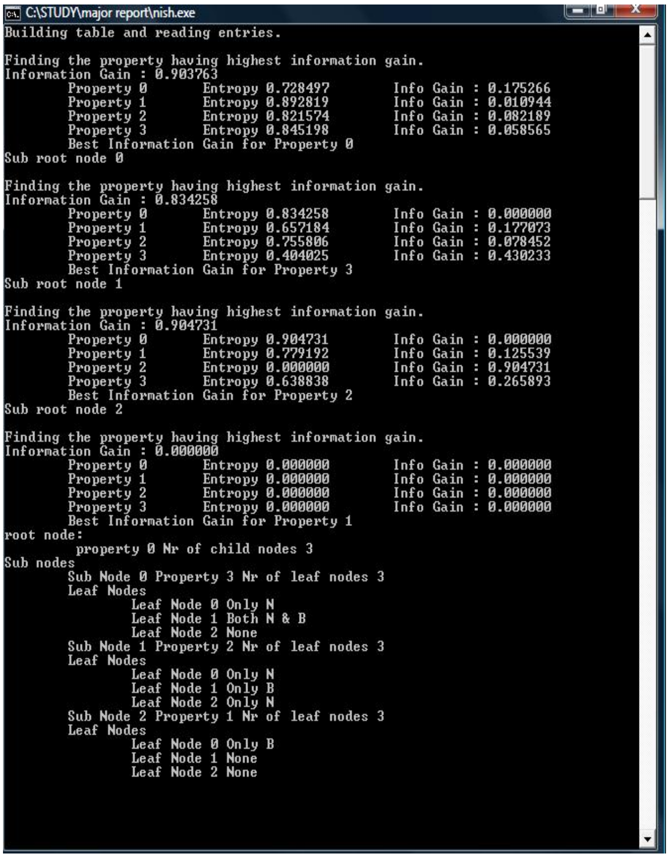


Fig. 3. Output in decision tree format.

F. Output in Command Prompt of ‘C’ Compilation



```
Building table and reading entries.

Finding the property having highest information gain.
Information Gain : 0.903763
        Property 0      Entropy 0.728497        Info Gain : 0.175266
        Property 1      Entropy 0.892819        Info Gain : 0.010944
        Property 2      Entropy 0.821574        Info Gain : 0.082189
        Property 3      Entropy 0.845198        Info Gain : 0.058565
        Best Information Gain for Property 0
Sub root node 0

Finding the property having highest information gain.
Information Gain : 0.834258
        Property 0      Entropy 0.834258        Info Gain : 0.000000
        Property 1      Entropy 0.657184        Info Gain : 0.177073
        Property 2      Entropy 0.755806        Info Gain : 0.078452
        Property 3      Entropy 0.404025        Info Gain : 0.430233
        Best Information Gain for Property 3
Sub root node 1

Finding the property having highest information gain.
Information Gain : 0.904731
        Property 0      Entropy 0.904731        Info Gain : 0.000000
        Property 1      Entropy 0.779192        Info Gain : 0.125539
        Property 2      Entropy 0.000000        Info Gain : 0.904731
        Property 3      Entropy 0.638838        Info Gain : 0.265893
        Best Information Gain for Property 2
Sub root node 2

Finding the property having highest information gain.
Information Gain : 0.000000
        Property 0      Entropy 0.000000        Info Gain : 0.000000
        Property 1      Entropy 0.000000        Info Gain : 0.000000
        Property 2      Entropy 0.000000        Info Gain : 0.000000
        Property 3      Entropy 0.000000        Info Gain : 0.000000
        Best Information Gain for Property 1
root node:
         property 0 Nr of child nodes 3
Sub nodes
        Sub Node 0 Property 3 Nr of leaf nodes 3
        Leaf Nodes
                Leaf Node 0 Only N
                Leaf Node 1 Both N & B
                Leaf Node 2 None
        Sub Node 1 Property 2 Nr of leaf nodes 3
        Leaf Nodes
                Leaf Node 0 Only N
                Leaf Node 1 Only B
                Leaf Node 2 Only N
        Sub Node 2 Property 1 Nr of leaf nodes 3
        Leaf Nodes
                Leaf Node 0 Only B
                Leaf Node 1 None
                Leaf Node 2 None
```

Fig. 4. Output in command prompt of ‘C’ compilation.

IV. CONCLUSION

Data mining as a technology can be used to analyze the customer data to find their exact need. This will help us to give more value to customer by increasing their information, also help in providing high quality services to them by understanding them. The decision tree tells what customers want the most. In this thesis ID3 algorithm is used, but modification is done. Instead of using Shannon Entropy, Havrda and Charvat Entropy has been used to find the information of different properties which is used as the node of decision tree. This modification has reduced the size of tree as well as decreased the rules, which will help to understand customer characteristics by which company growth and profit can be increased. I have proposed a new algorithm, i.e. instead of Shannon entropy we have used the Havrda and Charvat entropy ,as a result I can conclude that for lower value of alpha (α)=0.25, tree is small and less complex as compared to the use of Shannon entropy. As conclusion I can say that if we want to get less number of node and leaf in a tree and to make it more effective and less complex, we can use the Havrda and Charvat entropy instead of Shannon entropy and value of alpha (α) less than one will give decision tree with less number of nodes.

V. Scope for Future Work

In this research, the value of alpha (α) has been put as 0.25, but for further research we can take varying value of alpha (α) which may give different trees.

Various values of α have already been put in this study i.e. alpha (α) = 0.5, 0.75, 0.99, 5, 10, 100 which gave same tree as in the case of alpha (α) = 0.25. But, it has been observed that in case of α = 2, 3, 4 the tree turned out to be different with more number of leaf and high complexities.

Also, Instead of using Havrda and Charvat Entropy, Different Entropy can also be used for further research like Arimoto, Sharma-Mittal, Taneja, Sharma-Taneja, Ferreri, Sant'anna—Taneja ,Picard, Aczel-Daróczy.

References

[1] E. Thomas, "Data mining: definitions and decision tree examples," Stony Brook, State University Of Newyork.

[2] Data Mining Algorithms (Analysis Services - Data Mining). [Online]. Available:http://technet.microsoft.com/en-us/library/ms175595.aspx

[3] J. Su and H. Zhang, "A Fast Decision Tree Learning Algorithm," Faculty of Computer Science, University of New Brunswick, NB, Canada, E3B 5A3.

[4] W. Peng, J. Chen, and H. Zhou, "An Implementation of ID3 -- decision Tree Learning Algorithm," University of New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, Australia.

[5] C. F. L. Lima, F. M. de Assis, C. P. de Souza, "Decision Tree based on Shannon, R´enyi and Tsallis Entropies for Intrusion Tolerant Systems," Federal Institute of Maranh˜ao Maracan˜a Campus S˜ao Lu´ıs, MA – Brazil, Federal University of Campina Grande Campina Grande, PB – Brazil, Federal University of Para´ıba Jo˜ao Pessoa, PB – Brazil: Published in *The Fifth International Conference on Internet Monitoring and Protection*.

[6] J. Han and M. Kamber, "Data Mining: Concepts and Techniques (2nd edition)," Morgan Kaufmann Publishers, 2006

[7] J. R. Quinlan, "C4.5: Programs for Machine Learning," *Morgan Kaufmann Publishers*, Inc., 1993.

[8] M. Lee, Y. J. Kim, Y.-M. Kim, S. Cheong, and S. Song, "Classifying Bio-Chip Data using an Ant Colony System Algorithm," *International Journal of Engineering and Applied Sciences* vol. 2, no. 2, 2006

[9] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory* vol. 13, issue 1, pp. 21-27, January, 1967.

[10] L. Breiman, J. Friedman, L. Olshen, and J. Stone, "Classification and Regression trees. Wadsworth Statistics/Probability series," CRC press Boca Raton, Florida, USA, 1984.

[11] W. Peng, J. Chen, and H. Zhou. An Implementation of ID3 Decision Tree Learning Algorithm. [Online]. Available: web.arch.usyd.edu.au/wpeng/DecisionTree2.pdf

[12] T. Chen, B. C. Vemuri, A. Rangarajan, S. J. Eisenschenk, Group-Wise Point-Set Registration Using a Novel CDF-Based Havrda-Charvát Divergence.

[13] Q. Wang, Y. Wu, J. Xiao, and G. Pan, "The Applied Research Based on Decision Tree of Data Mining In Third-Party Logistics. Automation and Logistics," presented at 2007 IEEE International Conference on 08 October 2007, Jinan.

Mr. Nishant Mathur received M.Tech (Computer Science and Engineering) degree from Delhi College of Engineering.

Mr. Sumit Kumar received M.Tech (Computer Science and Engineering) degree from Indian Institute of Technology Guwahati then He joined Indian Institute of Technology Patna as Research Fellow in Department of Computer Science and Engineering.

Mr. Santosh Kumar received M.Tech (Computer Science and Engineering) degree from Indian Institute of Technology Guwahati then he joined Indian Institute of Technology Guwahati as Junior Project Fellow in Department of Computer Science and Engineering.

Mrs. Rajni Jindal is M.C.A., M.E.,SMIEEE , MWIE, LMISTE, LMCSI. Her specialized fields are Database Systems, Data Mining and Operating Systems. She published several research papers in reputed conferences & Journals. She is Assistant Professor at Delhi Technological University (Formerly Delhi College of Engineering).